‘Qualitative’ research, systematic reviews, and evidence-informed policy and practice

Angela Harden

Institute of Education, University of London

Thesis presented to the University of London for the Degree of Doctor of Philosophy

April 2007

Abstract

This thesis makes a distinctive contribution to debates about how to include and quality assess 'qualitative' research in systematic reviews. It analyses sets of quality criteria, assesses the impact of study quality on findings and compares 'quantitative' and 'qualitative' perspectives on quality. The research consists of a review of the literature and three new methodological studies. The first study surveyed and evaluated quality assessment tools, the second analysed the development of a new tool, and the third examined the relationship between the quality of 'qualitative' research and the findings of systematic reviews.

A large number of different quality criteria have been proposed for 'qualitative' research but assessment tools represent 'good practice' guides rather than aids to distinguish between 'good' and 'bad' studies. Continuous funding, a policy-focussed context, and a multi-disciplinary team which viewed research questions as drivers for quality assessment were important factors for developing a unique tool which did help to distinguish between studies. There was no straightforward relationship between study quality and the findings of reviews. However, excluding lower quality studies had little impact on review findings. Studies which made the biggest contribution to reviews were those with appropriate methods for the review question and findings displaying conceptual depth. In contrast to procedures for 'quantitative' research, engaging with study findings as well as study methods is important for assessing fully the quality of 'qualitative' research.

This thesis generates important empirical evidence for debates about how to assess the quality of 'qualitative' research. It shows how standard quality assessment protocols need to be altered better to fit 'qualitative' research, reveals how study quality can impact on review findings and demonstrates some problems with the terms 'qualitative' and 'quantitative'. Future debate in this area should focus on how to identify reliable answers to questions about intervention process, context and need.

I declare that the work presented in this thesis is my own work.

Word length

86,124

CONTENTS

Chapter 8 Discussion and conclusion

Acknowledgements

Thank you to Ann Oakley and Sandy Oliver, my PhD supervisors, who struck the perfect balance between providing critical guidance and letting me learn from my own mistakes. Ann's book, *Experiments in Knowing*, was an important inspiration for my thesis. There are many others who have helped me along the way. Thank you to all my colleagues and friends, past and present, at the Social Science Research Unit. In particular, I have benefited a great deal from discussions with Vicki Strange, James Thomas, Jonathan Shepherd, Ginny Brunton, Josephine Kavanagh, Rebecca Rees, Meg Wiggins, David Gough, Jo Garcia, Diana Elbourne, Chris Bonell, Kelly Dickson, Frances Kirk, and Katy Sutcliffe. Extra special thanks are due to my family Melissa Harden, Christopher and Edgar Thompson, and Helen, Don, Lauren and Danielle Souter. Special thanks also to Julie Jarvis, Sarah Marwood, Guy Genney, Graham Hill and Sue Sharpe. Above all, I would like to thank my mum and dad, Eileen and Owen Harden, for their continuous love and support - I am truly grateful.

This thesis is dedicated to the memory of my mum, Eileen Harden. She died before this thesis was complete, but she helped me through to the very end.

CHAPTER 1

Thesis aims and rationale, origins and programme of work, and key concepts, definitions and assumptions

1.1 Aims and rationale

In this thesis I aim to advance knowledge about how to include and quality assess 'qualitative' research in systematic reviews. My starting point is the demand from users of research evidence that systematic reviews - summaries which use rigorous and explicit methods to identify and integrate findings from multiple studies (Egger *et al.*, 2001; Petticrew and Roberts, 2006) - address issues of process, context, and need alongside effectiveness (Davies, 1999; Peersman *et al.*, 1999; Popay *et al.*, 1998). Although research evidence on the effects of interventions is crucial, policy-makers, practitioners, and researchers have argued that they also need to know whether interventions are acceptable to, and meet the needs of, their intended recipients; whether interventions are feasible to implement; and why interventions succeed or fail. Many different study types and data could provide answers to the kinds of questions posed by evidence users but in this thesis I focus on just two: i) process evaluations, which are designed to examine the way an intervention is implemented, delivered and/or received; and ii) studies of people's perspectives and experiences. These types of studies are often, but not always, 'qualitative' in nature.

'Qualitative' research has origins in the social science disciplines of anthropology and sociology (Denzin and Lincoln, 2005; Vidich and Lyman, 1994). As the work for this thesis unfolded, 'qualitative' research proved very difficult to define. What goes on under the label 'qualitative' research is actually very diverse, and features usually

highlighted as distinguishing 'qualitative' research from 'quantitative' research – for example, the use of words rather than numbers, the adoption of an inductive rather than a deductive approach, and the study of people and society in 'natural' rather than 'artificial' settings - can actually be features of either 'qualitative' *or* 'quantitative' research (Bryman, 1988; Hammersley, 1992; Oakley, 2000). In recognition of these problems, throughout this thesis I use single quotation marks around the word 'qualitative' (and 'quantitative') to achieve two things: firstly, to problematise the idea of 'qualitative' research as a coherent entity with universal meaning; secondly, to indicate that the studies of interest in this thesis – process evaluations and studies of people's perspectives and experiences – are defined primarily by their research question and can use methods usually associated with either 'qualitative' and 'quantitative' traditions.

In this thesis, I define 'qualitative' research as any type of research with one or more of the following features: a focus on context and meaning and the investigation of the world from the point of view of the people studied; the collection of textual data in the form of, for example, conversations, field notes, interview transcripts, photographs, or drawings; and the generation of concepts, explanations or theories through narrative interrogation of the data to identify patterns, themes, or contradictions. I include within my definition studies that may have used numbers or statistics in data collection and/or analysis. Although the latter are usually associated with 'quantitative' research, these methods of data collection and analysis can be, and are, used to investigate context and meaning and to examine the world from the point of view of the people studied. My definition, which draws on some of the common features of 'qualitative' research highlighted by other definitions (e.g. Hammersley, 1992; Miles and Huberman, 1994; Popay *et al.*, 1998), avoids the pitfalls of defining 'qualitative' research in opposition to 'quantitative' research and encompasses all of the studies of interest in this thesis.

Until recently, the focus of systematic reviews has been on 'quantitative' research to answer questions about the effects of interventions (Dixon-Woods *et al.*, 2001; Popay *et al.*, 2006). There is now growing enthusiasm from researchers, policymakers and practitioners for including 'qualitative' research in systematic reviews alongside 'quantitative' research, but methods for achieving this are underdeveloped (Dixon-Woods and Fitzpatrick, 2001; Mays *et al.*, 2005; Petticrew and Roberts, 2006; Popay, 2005). There is also an emerging genre of systematic reviewing activity that reviews 'qualitative' research in its own right, variously described as 'qualitative meta-synthesis' (Sandelowski *et al.*, 1987; Sandelowski and Barroso, 2003a; Thorne *et al.*, 2004), 'meta-study of qualitative research' (Paterson *et al.*, 2001; Thorne *et al.*, 2002), or 'meta-ethnography' (Britten *et al.*, 2002; Campbell *et al.*, 2003; Noblit and Hare, 1988). Regardless of whether the aim is to review 'qualitative' research alongside other types of research or on its own, there are similar uncertainties around how to search systematically for 'qualitative' research, how to bring together, integrate and synthesise its findings and - the topic of particular concern in this thesis - how to assess the quality of 'qualitative' research (Dixon-Woods *et al.*, 2006; Popay, 2005). One overarching debate around these uncertainties has been framed as whether the 'quantitative' model of systematic reviews will fit 'qualitative' research (e.g. Booth, 2001; Dixon-Woods *et al.*, 2006; Hammersley, 2001; Jones, 2004).

Quality in any type of research, whether 'qualitative' or 'quantitative', is a multi-dimensional concept (Furlong and Onacea, 2005; Miles and Huberman, 1994; Pawson *et al.*, 2003). One dimension, sometimes referred to as 'validity' or 'trustworthiness', refers to the extent to which research findings represent an accurate or truthful representation of the world (Cook and Campbell, 1979; Hammersley, 1992). This dimension is often judged according to the rigour of the methods used to conduct the research. Other dimensions of quality include whether

the research is useful and relevant, well reported and easy to read, or ethical in conduct (Furlong and Onacea, 2005; Miles and Huberman, 1994; Pawson *et al.*, 2003). In this thesis I focus mainly on the validity and trustworthiness dimension of quality. In practice, however, it is sometimes difficult to separate this dimension from others (Juni *et al.*, 2001a,b; Sandelowski and Barroso, 2002a).

Assessing the quality of 'qualitative' research has attracted much debate amongst social scientists who have argued over whether quality should be assessed using the same or different criteria to 'quantitative' research, who should assess quality, and whether quality can or should be assessed in relation to 'qualitative' research at all (e.g. Murphy *et al.*, 1998; Oakley, 2000; Seale, 1999; Spencer *et al.*, 2003). Those considering the role of 'qualitative' research in evidence-informed policy and practice and systematic reviews have taken up and extended these debates to argue over when quality assessment should take place in the systematic review process and whether studies should be included or excluded on the basis of quality (e.g. Attree and Milton, 2006; Dixon-Woods *et al.*, 2006; Sandelowski *et al.*, 1997). Despite the voluminous literature documenting these debates, theoretical and abstract exhortations dominate and there is very little empirical work to inform the debates. No-one is yet certain about the range of quality criteria that have been suggested and the extent to which these criteria differ or overlap. Similarly, there is uncertainty about which of the many suggested quality criteria are essential or useful. Moreover, it is still unclear whether the quality of 'qualitative' research included in systematic reviews makes any difference to review findings (Dixon-Woods *et al.*, 2006).

This thesis aims to make a new contribution to the literature by undertaking empirical analyses of the criteria proposed to assess the quality of 'qualitative' research and the relationship between study quality and research findings. The

underlying concern of the thesis is with the possibility that '*how* we know what we know' influences '*what* we know'. While this thesis examines the issue of quality in 'qualitative' research with particular reference to systematic reviews, it has the potential to contribute to debates about the quality of 'qualitative' research in general, and to broader debates about the nature and purpose of different research methods. This is because some of the issues involved in assessing the quality of 'qualitative' research are the same whether quality is assessed in a systematic review or for other purposes. Furthermore, attending to the quality of studies can illuminate issues in the design, and conduct of primary research, and the relationship between methods, findings and conclusions. Accordingly, although my primary aim is to make a methodological contribution with respect to systematic reviews, I also hope to contribute to the literature on research methods more generally.

The thesis has four specific aims:

1) To review the conceptual issues and methodological debates within the literature relevant to the topic of quality in both 'quantitative' and 'qualitative' research;
2) To identify, compare, and evaluate the quality criteria that have been proposed to assess the quality of 'qualitative' research;
3) To assess the relationship between the quality of 'qualitative' studies and the findings of systematic reviews that include them; and
4) To make recommendations for a) how to quality assess 'qualitative' research and how to include it in systematic reviews; b) further work to advance methods for assessing the quality of 'qualitative' research in systematic reviews and beyond; and c) primary research methods.

1.2 Thesis origins and programme of work

The programme of work for this thesis consists of a review of the literature on the topic of quality in research and three new methodological studies. In the literature review, I aimed to raise and describe the relevant conceptual and methodological debates in order to design new methodological work and seek clarity and context for the contribution of my thesis. The review is reported in two parts. The first part (**chapter 2**) focuses on 'quantitative' research and the second part (**chapter 3**) focuses on 'qualitative' research. Both chapters cover the literature on assessing the quality of research in general and the more specific literature on assessing the quality of research within a systematic review.

The findings of the literature review were used to inform the three new methodological studies. (The aims, design and methods of these studies are summarised in **chapter 4**.) The review revealed very little work with similar aims to this thesis for advancing knowledge about how to assess the quality of 'qualitative' research in systematic reviews. Unlike the literature for 'quantitative' research there have been few attempts to take a systematic approach to the identification, comparison and evaluation of existing criteria to appraise the quality of 'qualitative' research. The first and the second of the three new methodological studies were designed to fill this gap. The first study (**chapter 5**) aimed to identify, describe and compare sets of criteria proposed for assessing the quality of 'qualitative' research. These aims were achieved by conducting a survey and evaluation of quality assessment tools identified through an exhaustive and systematic search. The second study (**chapter 6**) analysed in depth the development of one particular tool for assessing the quality of 'qualitative' research. This tool was developed by myself and colleagues within the programmes of systematic review work in which this thesis originates. (These programmes and their relationship to this thesis are

discussed in full below.) The tool was of particular interest because it was so different from any of the tools surveyed in the first methodological study: it was embedded in an approach to quality assessment that was driven by the review question rather than methodology. This second methodological study aimed to identify the factors that influenced the development of this unique and innovative approach.

The review of concepts and methodological debates on the topic of quality in research reported in chapters 2 and 3 revealed that little attention has been paid in the literature to the possibility that the quality of 'qualitative' research in systematic reviews could make a difference to review findings. This is in contrast to work on 'quantitative' research, which has found that high quality trials and low quality trials give different answers about the effects of interventions (e.g. Kunz *et al.*, 2002). The third methodological study (**chapter 7**) was designed to address this gap and analysed the relationship between study quality and synthesis results in several systematic reviews that had included 'qualitative' research.

The above programme of work for my thesis originated in programmes of systematic review work conducted in the UK at the Evidence for Policy and Practice Information and Co-ordinating Centre (EPPI-Centre) based in the Social Science Research Unit (SSRU), Institute of Education, University of London. I am a member of the academic staff at SSRU and an associate director of the EPPI-Centre. I joined SSRU in 1997 to work in what was then called the Centre for the Evaluation of Health Promotion and Social Interventions (the EPI-Centre). I was attracted to the policy relevance of the work of SSRU with its focus on health, education, welfare and other services, and the relationships between professionals who deliver these services and the public who use them. Of particular interest was the systematic review work at the EPI-Centre. Although I had never heard of systematic reviews, I

was interested in the idea of bringing research together to provide readers with a reliable short cut to the evidence for particular policies and practices.

By the time I joined the team, SSRU and the EPI-Centre had completed several systematic review projects. This early work included a project to develop a database of controlled trials in education and social welfare funded by the Economic and Social Research Council (ESRC) and reviews on the effects of anti-smoking, reading recovery and juvenile delinquency programmes. With funding from various sources (e.g. the Medical Research Council (MRC), the (then) Health Education Authority and regional health authorities) the EPI-Centre undertook several more systematic reviews (e.g. workplace health promotion; prevention of accidental injury; sex education); expanded its infrastructure for conducting systematic reviews; ran training programmes for health promotion practitioners and policy makers in critical appraisal; and established the Health Promotion and Public Health Field of the Cochrane Collaboration (jointly with colleagues in Canada). A year after I joined the team the EPI-Centre won the first of what was to be successive rounds of three year funding from the English Department of Health (DH) for a programme of systematic review work in health promotion and public health (HP&PH). In 2000, with new funding from the (then) English Department for Education and Employment (DfEE) to establish a centre for evidence-informed policy and practice, the EPI-Centre became the Evidence for Policy and Practice Information and Co-ordinating Centre (EPPI-Centre).

The activities of the EPPI-Centre today fall into four main areas: 1) conducting, storing and communicating the results of systematic reviews; 2) co-ordination of national and international EPPI-Centre Review Groups which conduct systematic reviews; 3) building research capacity for systematic reviews; and 4) methodological development. The DH and the Department for Education and Skills (DfES) continue

to be major funders of EPPI-Centre work alongside a number of other UK and international government departments and organisations such as the English Department of Work and Pensions (DWP), the UK Social Care Institute of Excellence (SCIE), and the Canadian Council for Learning. The importance of the EPPI-Centre as a national resource for systematic reviews in the social sciences was recognised when the UK Economic and Social Research Council (ESRC) awarded us funds in 2005 to become the Methods for Research Synthesis Node of the National Centre for Research Methods, which aims to achieve a step change in social science methods and capacity.

I began the work for this thesis in 2001. At this time the two major programmes of work at the EPPI-Centre were in education (funded by the DfES) and HP&PH (funded by the DH). The focus of the DH programme was on producing systematic reviews to inform HP&PH policy concerns, whilst the focus of the DfES programme was on co-ordinating and building research capacity for systematic reviews amongst academics and other professionals working in education. I worked across both of these programmes, and one of the problems myself and colleagues had been trying to tackle was the question of how to include 'qualitative' research in systematic reviews. Within both the DH and the DfES work our remit was to develop methods for systematic reviews that went broader than 'just effectiveness' or 'just trials'. As noted at the very beginning of this chapter, users of evidence such as policy-makers and practitioners wanted systematic reviews to address issues of context, need, and process alongside effectiveness (Davies, 1999; Peersman *et al.*, 1999; Popay *et al.*, 1998). Including 'qualitative' research was also important for gaining credibility amongst some academics. 'Qualitative' methods were and still are very popular amongst UK social scientists and other academics working within fields such as education and HP&PH, and there is considerable resistance to using trials (Oakley, 1998; Oakley, 2006).

I was particularly curious about the inclusion of 'qualitative' research in systematic reviews. I had come to the EPPI-Centre with skills in both 'qualitative' and 'quantitative' research and enjoyed tackling the challenges of applying systematic review methods to process evaluations in a review of peer-delivered health promotion (Harden *et al.*, 1999c; Harden *et al.*, 2001a). Taking a decision to focus on the topic of including 'qualitative' research in systematic reviews for a PhD was therefore a logical step for me. When I originally started, my aim was to look at how to include 'qualitative' research at every stage of the systematic review process. However, as I began to engage with the literature, I decided to study how to include 'qualitative' research through a focus on the problem of how to assess its quality. Assessing the quality of 'qualitative' research and selecting high quality studies for review has been described in the literature as one of the most difficult and unresolved of all the challenges facing the systematic review of 'qualitative' research (Popay, 2005). Topic- and discipline-wise I kept my field of vision wide rather than focus on HP&PH or education. Having worked across public policy areas I saw that the underlying methodological issues for the inclusion of 'qualitative' research in systematic reviews were the same regardless of topic. I wanted my findings to be relevant across public policy fields and to make a contribution to the methodological literature in the social and health sciences.

Although this thesis originates in the programmes of work described above, and uses some of the completed systematic reviews in those programmes as sources of data, I undertook the review, and designed, implemented, analysed and wrote-up the three new methodological studies in this thesis. Although I developed the tool analysed in-depth in study 2 (chapter 6) with colleagues at the EPPI-Centre, I designed and carried out the analysis of the factors that influenced the tool development. Similarly, although I was only one of a large team who carried out the systematic reviews that were used as sources of data in study 3 (chapter 7), I

designed, implemented, and wrote-up the analysis of the relationship between study quality and synthesis results. Inevitability this thesis draws on the shared perspectives and ideas about research and research methods that are held by the team at the EPPI-Centre and SSRU. I have been involved in creating and articulating some of these ideas including question-led reviews, the value of including 'qualitative' research in systematic reviews for answering questions that go beyond effectiveness, and how to include 'qualitative' research in systematic reviews (e.g. Harden *et al.*, 2001a; 2004; Harden and Thomas, 2005; Harden, 2006; Oliver *et al.*, 2005; Thomas *et al.*, 2004). Other ideas on which I draw originated with my colleagues alone, and when this is the case I have applied and explored these ideas with full acknowledgement.

A number of assumptions about the nature and purpose of research and the nature of reality that research seeks to represent underpin my thesis. Implicit in what I have written so far is the assumption that research can and should be used to inform policy and practice and that the systematic review is a valuable method for bringing research together for this task. The rest of this chapter outlines these assumptions in more detail through a discussion and definition of each of the key concepts under study in this thesis: evidence-informed policy and practice; systematic reviews; 'qualitative' research; and research quality.

1.3 Evidence informed policy and practice

The context of this thesis is evidence-informed policy and practice (EIPP) and I use this term to refer to the collective set of activities and methods to make available and use the findings of research for making decisions about policy and practice. The use of 'evidence-*informed*' rather than 'evidence-*based*' is a deliberate choice and assumes a dynamic model of the relationship between research and policy and

practice. Within this relationship the role of research is 'illuminative' rather than 'definitive'; and the role of the policy-maker or practitioner is one of a translator or interpreter of research in a specific context for a particular purpose (Levacic and Glatter, 2001). This view also reflects the fact that research evidence will only be one of a number of resources drawn upon for making decisions (e.g. Davies, 1999; Oakley, 2000; Oliver and Peersman 2001; Weiss, 1979).

The fundamental principle lying behind EIPP is "collective uncertainty" about the effects of policy and practices in recognition that "professionals sometimes do more harm than good when they intervene in the lives of other people" (Chalmers, 2003, p24). Adopting EIPP means that policy-makers and practitioners will consider the results of research on whether interventions do more harm than good when they are choosing which interventions to implement in order to improve the lives of the public. Sackett *et al.* (1996, p71) define evidence-based medicine as the "conscientious, explicit and judicious use of current best evidence in making decisions about the care of individual patients". The words 'conscientious', 'explicit', and 'judicious' highlight that being evidence-based is not something that happens on an ad-hoc basis, but is an ongoing process that needs considerable skill and a willingness to be open about decision-making processes. Davies (1999, p108) defines evidence-based education as "a set of principles and practices for enhancing educational policy and practice" and describes its application as a process of locating judgement within the "available evidence....which explores and tests the professional experience of teachers, students and other constituents of learning communities" (Davies, 1999, p117).

These definitions highlight that EIPP not only offers potential in terms of improving policy and practice, but also an opportunity to challenge traditional notions of the 'expert' and to transform power relations between practitioners and the people they

work with. As Chalmers notes, "evidence of collective uncertainty about the effects of policies and practices should prompt professionals *and the public* to find out which opinions are likely to be correct" (Chalmers, 2003, p24 emphasis added). EIPP challenges policy-makers and practitioners to question the validity of their judgments and to use the best available evidence for their decision-making.

Of no less significance are the challenges from EIPP to the research community. Too much research of no practical relevance was a major criticism in a review of the organisation of healthcare research in the UK in the 1970s, which characterized universities as 'self-serving' with no input from research users (Black, 1997). Similar views have been expressed with respect to other areas of public policy in the 1990s. For example, in the annual lecture to the English Teacher Training Agency in 1996 – 'Teaching as a research-based profession' - David Hargreaves argued that educational research was offering "poor value for money" and that "the teaching profession has, I believe, been inadequately served by us [educational researchers]" (Hargreaves, 1996, p1). National and international reviews of educational research supported these arguments, concluding that research was having little impact on policy and practice because of its small-scale and non-cumulative nature and lack of accessibility to non-academic audiences (Hillage *et al.*, 1998; Organisation for Economic Co-operation and Development, 2003).

In the late 1990s there was much political support for a range of strategies to facilitate "evidence-based everything" (Oakley, 2002, p277) across government in the UK and internationally. EIPP was a cornerstone of the White paper '*Modernising Government*' published by the then newly elected Labour Government in the UK (Cabinet Office, 1999). In this document, ministers pledged to "improve our use of evidence and research" (p17) and the Secretary of State for Education at the time, challenged social scientists to produce and collate the evidence base:

> "We need to be able to rely on social science and social scientists to tell us what works and why and what types of policy initiatives are likely to be most effective. And we need better ways of ensuring that those who want this information can get it easily and quickly." (Blunkett, 2000, p21).

The EPPI-Centre is one example of significant investment in EIPP by government departments and research councils. As noted earlier, the EPPI-Centre houses national review facilities for conducting reviews and for building research capacity in research synthesis. Other UK examples have included the 'Evidence Network', established in 2000 and funded by the ESRC as a co-ordinating centre and network of nodes, which aim to bring social science research closer to the decision-making process; and the Centre for Reviews and Dissemination, set up in 1994 to review the effects of interventions in health and social care and funded by the English Department of Health (DH). Examples of international initiatives include the 'What Works Clearing House', funded by the US Department of Education and set up in 2002, which aims to provide a national source of reliable evidence on what works in education; the Cochrane Collaboration, set up in 1992 to collate, critically appraise, synthesise and disseminate evidence on the effectiveness of healthcare interventions; and the Campbell Collaboration, set up in 1999 to prepare, maintain and promote the accessibility of systematic reviews on the effects of interventions in education, crime and social welfare.

1.4 Systematic reviews

Reviews of research can serve many different purposes. They can help us work out what has already been done and what needs to be done; develop a new argument or perspective on a topic; identify the types of methods and theories that have been applied in a field; shed light on contradictory findings from similar studies; assess

generalisability and consistency of relationships across studies; synthesise evidence about the effects of interventions; provide a short-cut to the research literature on a particular topic; bridge the gap between research, policy and practice; and avoid duplication of effort and the considerable costs of embarking on new studies that may not be needed (Cooper and Hedges, 1994; Hart, 1998; Light and Pillemer, 1984; Mulrow, 1994; Petticrew and Roberts, 2006). As Gough and Elbourne (2002) have noted, literature reviews are considered by some to be a useful strategy in the accumulation of knowledge (e.g. Cooper and Hedges, 1994; Light and Pillemer, 1984), whilst others see reviews as a way to 'recast' the literature by analysing how research is historically and socially located within particular (dominant) conceptual frameworks (e.g. Lather, 1999; Livingstone, 1999).

This thesis is concerned with one type of review called a systematic review. A defining feature of a systematic review is the application of the scientific method to uncover and minimise bias and error in the selection and treatment of studies (Chalmers *et al.*, 2002; EPPI-Centre, 2006; Mulrow, 1994; Petticrew and Roberts, 2006). Concepts of 'bias' and 'error' are central to systematic reviews as these may lead to distorted or erroneous results. In research, 'bias' usually refers to any kind of systematic error introduced into research procedures - for example, when research is either consciously or unconsciously designed in ways that support prior beliefs - whereas 'error' refers to random mistakes (e.g. failure of tape recorder when collecting data in an interview study) (Hammersely and Gomm, 1997; Juni *et al.*, 2001a,b; Peersman *et al.*, 2001; Shadish *et al.*, 2002; Wallace, 1971). In systematic reviews, attempts are made to minimise the introduction of bias and error into the review process by, for example, identifying as much as possible of all the relevant research to avoid a selective sample of studies, or using standardised data collection protocols so that each study in the review is treated in the same way (Higgins and Green, 2006; Stock, 1994; White, 1994).

Attempts are also made to uncover bias and error in the individual studies included in the review. Studies are quality assessed and those of lower quality are sometimes excluded or given less weight (Higgins and Green, 2006; Juni *et al.*, 2001a,b; Wortman, 1994). Systematic review guidelines and handbooks often outline a set of discrete steps and processes to follow as an ideal type or standard for a reviewer to strive for (e.g. Campbell Collaboration, 2001; Cooper and Hedges, 1994; EPPI-Centre, 2006; Higgins and Green, 2006; Kahn *et al.*, 2001; Petticrew and Roberts, 2006). These steps and processes can be specified in advance in a review protocol which states what is to be done and why. Ideally, both the review protocol and the final review report are subjected to scientific peer review and made available for public scrutiny. Some systematic review guidelines recommend methods to protect against or uncover bias at the very beginning of the review process when deciding upon the question or hypothesis for review (e.g. EPPI-Centre, 2006; Jackson and Waters, 2005; Jackson *et al.*, 2005). Making review methods explicit and transparent can facilitate accountability and debate, replication, and review updates (Gough and Elbourne, 2002).

The final major stage in a systematic review is the synthesis, whereby the findings of the included studies are integrated to answer the review question. In comparison to other stages of the systematic review such as searching and quality assessment, defining what synthesis is and describing the processes involved is very difficult. Synthesis is defined in the *Oxford English Dictionary* as 'the process or result of building up separate elements, especially ideas, into a connected whole, especially a theory or system'. In research synthesis, 'ideas' are usually the findings of research, and the 'building up' of findings from individual studies into a 'connected whole' in a systematic review should ideally be guided by rigorous and explicit procedures. Statistical meta-analysis is one such rigorous and explicit method of synthesis that can be used when findings from studies are in a numerical form

(Deeks *et al.*, 2001; Lipsey and Wilson, 2001). In reviews which ask questions about the effects of interventions, the appropriate use of statistical meta-analysis - which is used to 'pool' the effect sizes from individual trials – is able to estimate the average balance of benefit and harm from an intervention with greater power and precision than less formal synthesis techniques such as simple 'vote-counting' (Bushman, 1994).

Alternative methods for synthesis which do not rely on statistical summaries are underdeveloped, although there is currently much on-going work to address this (e.g. Popay *et al.*, 2006; Campbell *et al.*, 2003; Dixon-Woods *et al.*, 2006; Pawson, 2006a; Sandelowski and Barroso, 2007; Thomas *et al.*, 2004). For example, Campbell *et al.* (2003) have applied and evaluated 'meta-ethnography' – a method developed in the 1980s for synthesising ethnographic studies in education (Noblit and Hare, 1988) – to synthesise the textual findings from 'qualitative' studies about illness experiences and health care, and Popay *et al.* (2006, p5) have produced guidance on 'narrative synthesis', a textual approach to synthesis which "tells the story of the findings of included studies". Although narrative synthesis is often used in systematic reviews, it is viewed with much suspicion as the processes used to produce a narrative synthesis are rarely stated or tested. The guidance produced by Popay and colleagues aims to help reviewers to introduce rigour and transparency into their narrative syntheses. This on-going development reveals that although systematic reviews have a long history (Chalmers *et al.*, 2002; Oakley, 2000; Rosenthal, 1991), methods are still evolving.

Despite their long history, systematic reviews have generally been the exception rather than the rule when one looks at the kinds of reviews that exist in a particular field (Chalmers *et al.*, 2002; Mulrow, 1987; Peersman *et al.*, 1999; Pillemer, 1984). For example, Peersman *et al.* (1999) surveyed 398 reviews of research on the

effects of health promotion and public health interventions and found that only a around a third had attempted to review the literature in a systematic way. When different reviews on the same topic or intervention reveal different findings about the effects of interventions (e.g. Oliver *et al.*, 1999a), examining whether reviews have been carried out in a systematic way can help to resolve the dilemma over which findings to rely on.

Chalmers and Haynes (1994, p363) predicted that "a substantial proportion of current notions about the effects of healthcare will be changed" when the research community "synthesises existing evidence thoroughly" by using systematic reviews. There is now a substantial body of evidence that supports these predictions in healthcare and beyond. These include interventions thought to be useful but later exposed in systematic reviews as harmful such as human albumin for the emergency treatment of burns and shock (Roberts, 2000) and the 'scared straight' intervention which increased crime among young people (Peterosino *et al.*, 2003); interventions thought to be useless but later exposed as beneficial such as corticosteroids to prevent complications from premature birth (Crowley, 1996); and interventions thought to be useful which have yet to be demonstrated as such when the evidence for them is reviewed in a systematic way such as sex education for the prevention of teenage pregnancy (DiCenso *et al.*, 2002).

1.5 Critiques of EIPP and systematic reviews

So far, I have suggested that evidence-informed policy and practice is valuable because it provides a framework to enhance the availability of research evidence for decision-making about the best solutions for health, educational, social, and economic problems. I have also suggested that systematic reviews offer advantages over traditional literature reviews about the effects of interventions

because they use explicit methods to try to minimise bias and error in their findings. However, EIPP and systematic reviews have come in for some serious criticism, especially from social scientists. Systematic reviews have been characterised by some social scientists as 'mechanistic', 'simplistic', 'atheoretical', 'naïve', 'outdated', 'stupid', 'boring', and 'positivist' (e.g. Evans and Benefield, 2001; Hammersley, 2001; MacLure, 2005; Schwandt, 1998). Gough and Elbourne (2002) have unpacked the ideological critiques underpinning these characterisations and identify several areas of concern that some of their fellow social scientists have expressed. These areas of concern suggest that EIPP and systematic reviews: wrongly assume a rational model of the relationship between policy and practice; are mechanisms for political control of the research agenda and professional practice; and marginalize 'qualitative' research. These are important issues to be debated.

The rational model of the relationship between research and policy assumes that research will drive policy, and this perspective has been criticised because it ignores the many other influences on the policy-making process (Davies, 1999; Nutley *et al.*, 2003; Weiss, 1979). However, as Gough and Elbourne (2002) note, EIPP and systematic reviews do not necessarily assume this model of the relationship between research and policy. As noted earlier, I see research evidence as illuminative rather then definitive in any policy or practice decision-making process, and research evidence as only one of a number of sources to draw on for making decisions. Fears about political control over the research agenda stem from calls for research users to be more involved than they have been in the past in setting research questions (Hammersley, 2002; Vulliamy and Webb, 2001). Because there has been significant UK government investment in initiatives to promote EIPP and systematic reviews, some academics are concerned that government will have too much control over the research agenda and that funding for 'blue skies' research – research that has no obvious real world application - will stop. However, systematic

reviews can, and often do, build in explicit steps for a range of user groups (including researchers) to be involved in deciding upon the question or topic to be researched (Garcia *et al.*, 2006; Jackson and Waters, 2005; Oliver, 2001; Thomas and Harden, 2003).

Social scientists have also criticised EIPP and systematic reviews because of their focus on 'quantitative' research and answering questions about 'what works?'. The concern that 'qualitative' research is marginalised in EIPP and systematic reviews is the one most relevant to this thesis, and there is evidence to support this concern. For example, the manuals for reviewers produced by the international Cochrane and Campbell Collaborations make no reference to 'qualitative' research (Campbell Collaboration, 2001; Higgins and Green, 2006). There are two main reasons why a focus on 'quantitative' research is thought to be unacceptable. For some, 'quantitative' methods, especially randomised controlled trials, are considered to be too simplistic for the study of the social world (e.g. Evans and Benefield, 2001; Hammersley, 2001; Jones, 2004). Others, including me, argue that a sole focus on 'quantitative' research is unacceptable because a) 'qualitative' as well as 'quantitative' research can help in the interpretation of systematic reviews of 'quantitative' evidence on the effectiveness of interventions; and b) systematic reviews should be used to address questions that go beyond effectiveness, and require 'qualitative' as well as 'quantitative' research for their answers (e.g. Dixon-Woods *et al.*, 2001; Petticrew and Roberts, 2006; Popay, 2006; Thomas *et al.*, 2004).

Questions which go beyond effectiveness include those about the processes involved in the implementations of interventions, and those about how people perceive and experience health, educational, social and economic problems (Davies, 1999; Oakley and Oliver, 2001; Popay, 2006). While it could be argued that

both 'quantitative' and 'qualitative' research is needed to fully address these questions, many of those who discuss the relevance of 'qualitative' research for policy and practice have emphasised its strengths for illuminating issues of process, perspective and experience (e.g. Fitzpatrick and Boulton, 1994; Giancomini and Cook, 2000a; Patton, 1990; Popay *et al.*, 1998; Reichardt and Cook, 1979). For example, Reichardt and Cook (1979) argue that 'qualitative' research can help us to monitor the implementation of an intervention, describe the contextual factors involved, elicit feedback from intervention providers and recipients, and generate explanations for how an intervention achieved (or did not achieve) the outcomes it intended to. Similarly, Patton (1990) argues that 'qualitative' methods are particularly well suited to studying issues of process in intervention evaluation as "depicting process requires detailed description; the experience of process typically varies for different people; process is fluid and dynamic; and participants' perceptions are a key process consideration" (p95). Beyond interventions, Popay *et al.* (1998) argue that 'qualitative' research can help us to identify and understand the factors that people consider to be important influences in their lives. This, in turn, can help us to explain why people behave in the ways that they do, which has relevance for developing interventions to test in the future.

1.6 Defining 'qualitative' research

I have made a case for why it might be valuable to include 'qualitative' research in systematic reviews on the basis of what it might enable us to achieve. However, as already noted in the opening section of this chapter, 'qualitative' research is extremely difficult to define. The terms 'qualitative' and 'quantitative' have come to mean much more than the methods used to conduct research. They are often used to refer to separate and competing 'paradigms' with different epistemological, philosophical and ethical underpinnings. Bryman (1988) observes that the

emergence of 'qualitative' and 'quantitative' as competing paradigms took root in the 1960s:

> "Increasingly, the terms 'quantitative research' and 'qualitative research' came to signify much more than ways of gathering data; they came to denote divergent assumptions about the nature and purposes of research in the social sciences" (Bryman, 1988, p3)

The competing paradigm perspective helps to locate the concerns expressed by some social scientists about the 'quantitative' focus of EIPP and systematic reviews. 'Quantitative' research is often characterised by those adopting the 'qualitative' paradigm as based on an outdated and discredited 'positivist' model of natural science which, it is argued, is inappropriate for the study of the social world. For example, in the 1980s educational researchers Egon Guba and Yvonna Lincoln advocated the 'naturalistic paradigm' for producing knowledge about the social world, defined as "dedicated to the study of behavioural phenomena *in situ'* (Guba and Lincoln, 1981, pxi). Although they did not rule out the use of 'quantitative' methods within such a paradigm, they argued strongly that the most appropriate methods for any research involving human behaviour should be 'qualitative' rather than 'experimental' (Guba and Lincoln, 1981).

'Quantitative' research is also characterised as 'unethical' and 'insensitive'. These charges against quantification have been made by authors from many different areas, but are particularly associated with feminism (Denzin and Lincoln, 2005; Reinharz, 1984), health promotion (Davies and McDonald, 1998: Tones and Tilford, 1994) and nursing (Bonell, 1999). Oakley (2000) argues that such associations are extremely significant. She traces the historical trajectories of 'qualitative' and 'quantitative'/experimental ways of knowing in order to expose the ideological

representations of each which are constructed in the competing paradigm view. She demonstrates how such representations are intimately bound up with issues of gender and power, and lead those working within disciplines on the fringes of the mainstream to reject 'quantitative' ways of knowing in favour of 'qualitative':

> "..the historical development of ways of knowing gave both 'qualitative' and 'quantitative' research sets of ideological associations which effectively linked the former to the interests of the underdog – to those whose position excludes them from mainstream power and authority, motivating them to construct knowledge from below, rather than above" (p296)

Much has been written on deconstructing the divide between the 'qualitative' and 'quantitative' paradigms and dispelling the misconceptions about each which have arisen as a result of the divide (e.g. Bryman, 1988; Cook and Reichardt, 1979; Hammersley, 1992; Oakley, 2000). These authors argue that features which have been aligned with either 'qualitative' or 'quantitative' can actually be features of *either* type, and that it is relatively easy to find examples of studies to demonstrate this. For example, 'qualitative' and 'quantitative' research can both be 'inductive' or 'deductive'; use numbers or words; test or generate hypotheses or theories; use methods that are ethical or unethical; study phenomenon in 'natural' or 'artificial' settings and so on. Furthermore, what is often taken as the 'unifying' or 'central motif' of qualitative research – a concern with understanding the social world from the point of view of the actors within it (e.g. Bryman, 1988, p8; Denzin and Lincoln, 2005, pxvi) – is actually the concern of many 'quantitative' researchers too (Hammersley, 1992).

Oakley (2000, p293) is particularly interested in the relationship between experimental 'ways of knowing' and 'qualitative' ones. She argues that although

experimental methods are characterised in the paradigm wars as the epitome of everything that is wrong with quantification (and that those who conduct experimental research "necessarily commit atrocities of one sort or another"), 'experimental' ways of knowing are actually part of our 'ordinary qualitative world' and that 'qualitative' and 'experimental' ways of knowing are "thus fused at ground level, in the everyday experiences from which all knowledge comes".

Indeed Donald Campbell, a strong advocate for experimentation in the social sciences, did not necessarily associate experimental methods with the usual 'hallmarks' of quantitative research:

> "Good experimental design is separable from the use of statistical tests of significance. It is the art of achieving interpretable comparisons and as such would be required even if the end product were to be grouped percentages, parallel prose case studies, photographs of groups in action etc." (Campbell and Stanley, 1966, p22)

Reichardt and Cook (1979) suggest that 'quantitative' understandings always need to build on 'qualitative' understandings. They define 'quantitative' methods as "techniques of counting, scaling and abstract reasoning" which need to build on 'qualitative' methods which they define as "techniques of personal understanding, common sense, and introspection" (p22). In other words, for 'counts' or 'scales' to be meaningful, they must be based on an understanding of what is meaningful to the people being studied.

What all of the above serves to demonstrate is that we cannot rely on conventional depictions of what 'qualitative' research is (or does) and what 'quantitative' research is (or does). Although we might all 'know' what kind of research is denoted by the labels 'qualitative' or 'quantitative', definitions which can withstand the diversity of

what actually goes on under these labels are elusive. Indeed, the authors previously cited who have shown that many of the assumed differences between them are, on closer scrutiny, illusory, have argued that we should dispense with the labels, or at least try to get beyond all the 'imagined' differences between them which can be unhelpful in our thinking. However, the terms (and the 'baggage' that go with them) are pervasive. One of the consequences of the 'paradigm wars' is that, as David Silverman (2001, p25) notes, if definitions of 'qualitative' research are given in opposition to 'quantitative' research, it is difficult to avoid assuming "a fixed preference or pre-defined evaluation of what is 'good' (i.e. 'qualitative') and 'bad' (i.e. 'quantitative') research. Although Silverman (2001, p32) is reluctant to fall into this trap, even he does not escape it when he suggests that "The methods used by qualitative researchers exemplify a common belief that they can provide a 'deeper' understanding of social phenomena than would be obtained from purely quantitative data'" and that "a dependence on purely quantitative methods may neglect the social and cultural construction of the 'variables' which quantitative research seeks to correlate" (p40).

Bryman (1988) favours a 'technical' rather than an 'epistemological' approach to defining 'qualitative' and 'qualitative' research. He argues that focusing on the technical aspects of a decision to use one type of research over the other will facilitate "an appreciation of the common technical problems faced by practitioners working within the two traditions" (p12) and will help researchers to question whether the two methods are really as different as the 'epistemological' approaches suggest. With this 'technical approach', Bryman goes on to outline the differences between approaches in terms of levels of precision:

> "Social surveys are ....likely to be preferred when .... there is concern to establish cause-and-effect relationships. Experiments are even stronger in this department

....Qualitative researchers are not uninterested in causes, in that they are frequently concerned to establish how flows of events connect and mesh with each other in the social contexts they investigate, or how their subjects perceive the connections between facets of their environment. However, survey and experimental researchers tend to be much more concerned with the *precise delineation of a causal factor, relative to other potential causes*" (Bryman, 1988, p110, emphasis added)

Like Bryman, Hammersley (1992, p163) sees 'precision' as crucial for the distinction between 'qualitative' and 'quantitative'. He argues that when choosing between methods we are faced with a range from more to less precise data, and questions whether 'more precise data' should always be desirable. Decisions about appropriate levels of precision should then "depend on the nature of what we are trying to describe, on the likely accuracy of our descriptions, on our purposes, and on the resources available to us, not on ideological commitment to one methodological paradigm or another". Although Hammersley gives some indication of when it might be appropriate to choose less precise data (when a wide focus is needed), it is frustrating that he does not go on to elaborate any further. In fact, in a later footnote, Hammersley argues that 'more precise data' may not be appropriate within the social sciences when he argues that "there are severe practical limits to the level of combined precision and accuracy that can be achieved in many areas of social science" (Hammersley, 1992, p173).

In various forms, Bryman, Hammersley and Silverman hint at the value of thinking in terms of research questions first and then matching the research method accordingly, rather than vice versa. A purpose-driven or question-led approach provides us with a way to 'soften' the polarisation of 'quantitative' and 'qualitative' methods because the issue is not whether 'qualitative' research is better that 'quantitative' research *per se*, but about which method is best under which

circumstances. A softening of the polarisation between 'qualitative' and 'quantitative' methods has been identified by Oakley (2002, p281), as a key issue for the social sciences to address in "accommodating itself to the challenge of the evidence movement."

As noted at the beginning of this chapter, I use single quotation marks around the word 'qualitative' (and the word 'quantitative') to highlight the problems in defining these terms and to raise the possibility that 'qualitative' research does not exist as a coherent and universal entity. In this thesis I adopt a question-led approach to 'qualitative' research rather than an epistemological or a technological approach. A question-led approach is in line with my practical purpose to respond to questions about need, context and process in the development and evaluation of interventions. Unlike some definitions of 'qualitative' research (e.g. Denzin and Lincoln, 2005), I do not equate 'qualitative' research with 'critical' or 'post-positivist' approaches to inquiry because 'qualitative' research can be carried out within 'mainstream' or 'positivist' approaches to inquiry too (Seale, 2004; Willig, 2001).

For the purposes of this thesis, I consider as 'qualitative' any type of research with a focus on context and meaning and the investigation of the world from the point of view of the people studied. This definition avoids the pitfalls of defining 'qualitative' research in opposition to 'quantitative' research, but because it is very inclusive – it covers studies that use numbers and statistical analysis as well as those that collect textual data and use narrative methods of analysis - it is not a very useful descriptor. I therefore attempt to be very precise about the methods used when discussing the individual studies under examination in this thesis. Rather than describe studies as 'qualitative', I spell out study aims and conceptual framework, data collection methods and data analysis methods. As will be discussed later, taking such a

precise approach was essential to the progression of the analyses in the new methodological studies that I carried out.

1.7 Quality in research

Discussions about quality are 'high stakes' for the research community. The status of research as a legitimate and useful activity depends on quality as well as individual researcher reputations and their collective status as a professional group (Hargreaves, 1996; Oakley, 1998). Assessing quality is therefore a 'risky business' and the evidence-informed policy and practice movement, along with its key tool the systematic review, has brought such a business to the fore. Of course, the research community has always been concerned with issues of quality. This concern is evident in the explicit systems which have been devised to ensure quality (e.g. peer review, 'core' elements in training programmes, ethical codes, guidelines for research reporting) and the norms and principles of working within a scientific or social scientific research community (e.g. a culture of reflection, criticism and debate). The evidence-informed policy and practice movement, however, calls into account these very systems of 'quality control', as well as the quality of research itself.

Quality in research is a multi-dimensional concept. Two recent frameworks for quality standards in research propose between four and six dimensions (Pawson *et al.*, 2003; Furlong and Oancea, 2005). Pawson *et al.* (2003) include six dimensions of quality: *transparency* - whether the process of knowledge creation is open to outside scrutiny; *accuracy* - whether knowledge claims are supported by the data or information on which they are based; *purposivity* - whether the methods used in the research were suitable for the 'task in hand'; *utility* - whether the knowledge generated is a useful answer for the question posed; *propriety* – whether knowledge

has been created with due respect for legal and ethical matters; and *accessibility* - whether the knowledge has been presented in a way that meets the needs of the knowledge user. Furlong and Oancea (2005) propose four dimensions which have much overlap with those suggested by Pawson *et al.* (2003): *epistemic* – whether the research is trustworthy, makes a contribution to knowledge, is transparent, and ethical; *technological* – whether the research is salient and timely, fit for purpose, and responsive to the needs of research users; *capacity building and value for people* – whether the research has involved partnership, collaboration and engagement with users, is plausible from a practitioners perspective, and stimulates personal growth; and *economic* – whether the research is, for example, cost-effective and competitive.

Engaging in any type of review of research offers the opportunity to examine issues of quality in research in detail. Systematic reviews, by their very definition, *demand* such a detailed examination. For example, Egger and Davey-Smith (2001, p23) list quality assessment as one of a number of distinctive features of systematic reviews of trials: "The formulation of the review question, the *a priori* definition of eligibility criteria for trials to be included, a comprehensive search for such trials *and an assessment of their methodological quality*, are central to high quality reviews" (emphasis added). Similarly, in the stages of research synthesis proposed by Cooper and Hedges (1994, p8), "applying criteria to separate 'valid' from 'invalid' studies" is seen as the primary function of their 'data evaluation' stage. However, this detailed attention to quality does not cover all of the quality dimensions just discussed above. Because systematic reviews aim to assess the quality of studies in order to produce a reliable answer to the question under study, the aspect of research quality under scrutiny in systematic reviews lies within the epistemic dimension of the framework from Furlong and Oancea (2005), and within the accuracy dimension of the framework from Pawson *et al.* (2003).

I also focus on epistemic and accuracy issues within this thesis because I am concerned with how to assess the trustworthiness of 'qualitative' research in relation to questions about intervention processes and people's perspectives and experiences. This concern is underpinned by particular ontological and epistemological positions. I assume that there is a 'reality' (e.g. interventions, poverty, workplaces) that exists independent of us as researchers about which we can have knowledge, but I assume that such knowledge can only ever be a representation of reality (Mark, 2000). This is because the process of knowledge generation is a social one and knowledge is the product of both 'reality' and our ways of knowing about that reality (Bailey, 2001). Although I assume that knowledge can only ever be a representation, I do believe that these representations can and should be judged in terms of their likely truth and that it is meaningful to trust particular pieces of research more than others. I therefore disagree with those who adopt a 'relativist' epistemology and consider all research to produce equally valid but different versions of the 'truth' (e.g. Smith, 1984).

I also disagree with those who argue that we should abandon the quest for validity and truth and instead judge the worth of research solely in terms of whether it serves the interests of oppressed groups (e.g. Lincoln, 1995) or according to its 'fertility' for generating new ideas or solutions (e.g. Lather, 1993). I agree with Hammersley (1992, p66) who argues that if we abandon attempts to establish validity, there would be little justification for research activity. Unlike Hammersley, however, I believe that the 'business' of establishing validity should not just be the concern of the academic community. I agree with Oakley (2000, p4) that "the goal of would-be knowers is the elimination of as much bias or distortion as is possible in what it is that counts as knowledge. This means meticulous, systematic, transparent, sensitive striving for descriptions of 'reality' that satisfy not primarily knowers' needs for professional and scientific recognition, but the much more

generous task of helping human beings to make informed decisions about how best to lead their lives".

1.8 Summary

In this chapter I have described the aims and origins of my thesis, discussed issues surrounding the definitions of the major concepts involved, and outlined some of my assumptions about the nature of 'qualitative' research, systematic reviews, and evidence-informed policy and practice. In particular, I have discussed the difficulties involved in defining 'qualitative' research, the value of reviewing and synthesising research in a systematic way, and the importance of assessing the quality of research in systematic reviews and beyond. I have also highlighted how my thesis - which aims to advance knowledge about how to include and quality assess 'qualitative' research in systematic reviews - has the potential to contribute to fundamental debates about research methods and the assessment of quality in research, despite its origins in practical questions about how systematic reviews can address need, context and process, as well as effectiveness.

The next two chapters report a more detailed discussion and review of the literature relating to the topic of quality in research. Whilst this chapter has located my thesis within debates about the use of research to inform policy and practice and the inclusion of 'qualitative' research within systematic reviews and EIPP, the next two chapters locate my thesis within specific debates about quality in research. As noted earlier, the review is reported in two parts. The first part (**chapter 2**) focuses on 'quantitative' research and the second part (**chapter 3**) focuses on 'qualitative' research.

CHAPTER 2

A review of the conceptual and methodological literature relating to the topic of quality in 'quantitative' research

2.1 Aims and rationale

In the review reported in this chapter, I aim to describe and discuss some of the major themes in the literature relating to quality in 'quantitative' research. Despite my focus in this thesis on 'qualitative' research, it was important for me to review the literature on quality in 'quantitative' research for two reasons. Firstly, I found that discussions in the literature on quality in 'qualitative' research nearly always revolved around whether 'qualitative' research can be assessed in the same way as 'quantitative' research (see **chapter 3**). It was therefore important to review the 'quantitative' literature to fully understand the debates in the 'qualitative' literature. Secondly, when I started the work for this thesis in 2001 very little had been written on the topic of including 'qualitative' research in systematic reviews. I therefore looked to the 'quantitative' literature as a starting point to generate ideas for new methodological work to advance knowledge about how to include and assess the quality of 'qualitative' research in systematic reviews.

The review reported in this chapter focuses on one type of 'quantitative' research that measures the impact on outcomes of interventions (e.g. 'randomised controlled trials', 'experiments' and 'quasi-experiments'). Although there is a literature on quality in other types of 'quantitative' research such as surveys (e.g. Biemer and Lyberg, 2003; deVaus, 2002; Hoinville and Jowell, 1978; Oppenheim, 1966), I focus on research testing the effect of particular policies and practices as examples of

'quantitative' research through which to explore the conceptual issues and methodological debates relating to quality. I chose this type of 'quantitative' research as an example because my aim was to cover quality assessment issues with particular reference to systematic reviews. Compared to survey research, there is much better coverage in the literature on assessing the quality of research in systematic reviews with respect to 'randomised controlled trials', 'experiments', and 'quasi-experiments'. To identify literature for this chapter, I trawled for relevant references in the following sources: three major systematic review handbooks (Cooper and Hedges, 1994; Egger *et al.*, 2001; Higgins and Green, 2006); a list of methodological reports published by the NHS Health Technology Assessment Programme; and the register of methodological studies in the Cochrane Library.

Like the focus of my thesis overall, this review concentrates on one particular dimension of quality, the dimension concerned with the accuracy or trustworthiness of research findings. In evaluations of the impact on outcomes of particular policies and practices, a 'quantitative' approach is used to measure the size of any intervention effect. Whether or not these effect sizes are accurate and trustworthy can be affected by a variety of issues such as how outcomes were measured (e.g. were they both reliable and valid?), the size of the sample used (e.g. was it big enough?) and the nature of the sample (e.g. was it the right kind of sample?). A crucial prior issue is, however, the design of the study that has been used to draw conclusions about whether the intervention under test is the cause of any observed changes in outcomes. Although researchers have used many different study designs (including 'qualitative' research) to examine cause and effect, 'experiments' or 'randomised controlled trials' are generally considered to be the strongest for drawing conclusions about cause and effect (Cook and Campbell, 1979; Kelijnen *et al.*, 1997; Oakley and Fullerton, 1996).

2.2 Randomised controlled trials, experiments and quasi-experiments

The use of experimental designs in both the social sciences and healthcare has a long history dating from at least the seventeenth century (Chalmers, 2001; Oakley, 2000). The following quote from the Flemish physician Jean Baptiste van Helmont in the seventeenth century cited by Chalmers (2001, p1157) and Oakley (2000, p148), illustrates the simple idea lying behind experimental design – to take a group of similar people (those who have 'fevers, pleurisies etc'); divide these into groups at random ('cast lots'); introduce a treatment into one group and a different or no treatment into the other ('blood letting and sensible evacuation'); and then examine differences in outcomes between the groups (in this case the number of funerals):

> "...come down to the contest ye Humorists: Let us take out of the Hospitals, out of the Camps or from elsewhere, 200, or 500 poor People, that have Fevers, Pleurisies, etc. Let us divide them in Halfes, let us cast lots, that one half of them may fall to my share and the other to yours; I will cure them without bloodletting and sensible evacuation; but do you do as ye know....we shall see how many Funerals both of us shall have..."

Chalmers (2001) notes that this description illustrates the fundamental goal of experimental design, to ensure that 'like is compared with like' in order to test fairly between alternative treatments or other kinds of interventions. As experimental designs became adopted within researchers' methodological tool-kits in the health and social sciences, each respective discipline adopted different sets of technical terms to describe them. Whilst the general term 'experimental design' is used most often within social science, the term 'randomised controlled trial' is the one adopted in the healthcare literature. The latter term highlights the defining feature of this particular experimental design - 'random allocation' of groups or individuals to intervention and control/comparison groups. Using random methods to allocate

means that it should not be predictable in advance which people will be assigned to each group and each person has an equal chance of being in either group. Random allocation is considered to be the most efficient means (both in the practical and statistical sense) of obtaining unbiased comparison groups so that when the intervention under test is introduced to one group it is, on balance, likely to be the only difference between them.

There are several variants of the randomised controlled trial (RCT) and a range of other experimental designs (Campbell and Stanley, 1966; Roberts and Sibbald, 1998; Torgeson and Roland, 1998). For example, there are 'single' and 'double-blind' randomised controlled trials in which attempts are made to 'hide' details of who is getting the intervention under test and who is not from the investigators and/or participants in the trial (Day and Altman, 2000); the 'pretest-posttest control group design' in which measures on the outcome of interest are taken before and after the introduction of the intervention under test (Campbell and Stanley, 1966); and the 'posttest only control group design' in which outcomes are only measured after the introduction of the intervention (Campbell and Stanley, 1966). There are also several 'quasi-experimental' designs that do not involve random allocation of individuals from the same population to control and experimental groups. For example, in a 'non-equivalent control group' design, the experimental and control groups are formed by "naturally assembled collectives such as classrooms" rather than random allocation (Campbell and Stanley, 1966, p47).

2.3 'Threats to validity' and 'sources of bias' in experiments

Researchers working within the health and social sciences have developed similar frameworks for helping us to design and assess the quality of research employing experimental designs. The 'threats to validity' framework was developed by the

social scientist Donald Campbell and his colleagues who were working on studies of the effects of educational interventions. Their validity framework, published in the 'Handbook of Research on Teaching' in 1963 (Campbell and Stanley, 1963) and later reproduced as a stand-alone text 'Experimental and Quasi-Experimental Designs for Research' (Campbell and Stanley, 1966), is said to have "revolutionised and simplified establishing the validity of a scientific study" (Elek-Fisk *et al.*, 2000, p46). The framework has undergone several revisions (e.g. Campbell, 1986; Cook and Campbell, 1979; Shadish *et al.*, 2002), but because the most substantive revisions appear in Cook and Campbell (1979), it is this source that is referred to below in the description of the framework.

For Campbell and colleagues, 'validity' is used to "refer to the best available approximation to the truth or falsity of propositions, including propositions about cause" (Cook and Campbell, 1979, p37). The central premise of the framework is a simple one, to enhance the validity of causal inferences the researcher must reduce the threat of, or rule out, as many plausible rival hypotheses for the observed relationship as possible. The framework outlines a range of threats to validity grouped under four inter-related types of validity: 'internal validity'; 'statistical conclusion validity'; 'external validity'; and 'construct validity'. Internal validity and statistical conclusion validity both refer to the extent to which a study has protected against drawing false negative or false positive causal inferences (i.e. concluding that there is an effect when there is none or concluding no effect when there is an effect). While statistical conclusion validity may be threatened by the inappropriate use or interpretation of statistical tests, internal validity may be threatened by plausible alternative explanations for any observed effect on outcomes beside the intervention. Cook and Campbell (1979) highlight the strengths of experimental designs over quasi or non-experimental studies in assuring internal validity because a larger number of validity threats can be ruled out by randomly allocating

individuals to intervention and control groups (e.g. effects due to history, maturation or repeat testing on outcome measures). They also acknowledge, however, the fact that other threats to validity are specifically introduced *because* of the employment of control or comparison groups (e.g. compensatory rivalry by participants receiving less desirable treatments). Construct validity and external validity are both concerned with abstracting more general conclusions from the results of a study. Construct validity refers to the validity with which one "can make generalisations about higher-order constructs from research operations" (Cook and Campbell, 1979, p38). In other words, assessing construct validity involves assessing how well outcomes and interventions have been defined, measured and/or implemented. External validity is defined as "the approximate validity with which we can infer that the presumed causal relationship can be generalised to .....and across different types of persons, settings, and times" (Cook and Campbell, 1979, p37).

The influence of Campbell and Stanley's validity framework has been wide-ranging. Not only has it underpinned approaches to the evaluation of educational and other social programmes (and approaches to systematic reviews of such evaluations), but it has also been the starting point for those articulating other validity frameworks for 'qualitative' or 'quantitative' case studies (the 'rival explanations' approach described by Yin, 2000) and for other types of non-experimental or 'qualitative' studies (e.g. LeCompte and Goetz, 1982). Indeed, the sociologist Howard Becker, who has applied the Campbellian framework to the use of photographs as data in social research, argues that in any type of research we need to:

> "decide whether a proposition is true (or perhaps, better, whether we ought to believe it) by thinking explicitly of all the reasons we might have to doubt it, and then seeing whether the available evidence requires us to take these doubts seriously. If the evidence suggests that we need not entertain these doubts, that these threats to the

validity of our idea are not sound, then we can accept the proposition as true" (Becker, 1979, p107)"

Campbell and Stanley's (1966) validity framework has also been influential in the health sciences. In this area, effort has primarily focused on threats to validity for the randomised controlled trial (RCT). Unlike the social sciences, there is a much greater consensus that an RCT is a) a feasible method of evaluation in many, but not all, situations; and b) when properly designed, executed and analysed, the RCT is the study design which has the greatest chance of producing unbiased assessments about the effect of interventions. Researchers in healthcare have focused on internal validity rather than external validity. Juni *et al.* (2001a) describe what they see as the four most important types of bias which threaten the internal validity of a trial. These can all arise when systematic differences between control/comparison groups and intervention groups are introduced by the design, execution or analysis of a trial.

Systematic differences between control/comparison groups and intervention groups can occur as a result of factors within four main categories (the name of the bias associated with each is given in brackets): i) differences between the people allocated to intervention and control/comparison groups on factors related to the outcome(s) under investigation (selection bias); ii) the preferential provision of interventions apart from the one under evaluation to comparison groups (performance bias); (iii) more favourable assessment of outcomes given to those in the intervention group or less favourable assessment of those in the control/ comparison group (detection bias); and (iv) the occurrence and handling of participant attrition which result in differences between the people remaining in the intervention group as compared to the control/comparison group (attrition bias).

Selection bias and attrition bias warrant a little more explanation. Selection bias can arise from inadequate allocation procedures. Allocation procedures should aim to "ensure as far as possible that, on average, the people who make up the comparison groups are comparable in respect of prognosis and responsiveness to treatment" (Kleijnen *et al.*, 1997, p95). Juni *et al.* (2001a) describe two threats to this: inadequate generation of the allocation procedures so that they are predictable in advance (which they suggest is more likely with quasi-random allocation procedures such as alternation); and inadequate concealment of allocation so that investigators of participants can see whether the next allocation is to the control or comparison group. Attrition bias can occur with the loss of participants after allocation, either because it is found on further investigation that they do not meet the eligibility criteria for the trial, or because they are unavailable at follow-up. Because it is likely that these participants will differ on factors related to outcomes from those remaining in the study it is important that any differences are examined. When possible, analysis of the trial should be done on all participants as allocated, not just on those remaining in the study.

It is clear that perspectives from researchers in the health and social sciences overlap considerably, although a focus on RCTs within the healthcare literature means that considerably more attention has been paid to identifying inadequate and adequate allocation procedures. This issue is barely touched upon in Cook and Campbell (1979) when they outline threats to validity. Regardless of whether one uses the terminology of 'sources of bias' or 'threats to validity' the key tasks for evaluating the effects of interventions are a) to identify relevant sources of bias or plausible rival hypotheses and b) to design, implement and analyse the study to reduce the likelihood of bias or the plausibility of rival hypotheses. The key task of quality assessment in 'quantitative' research is to assess whether a study *has* reduced the likelihood of bias or the plausibility of rival hypotheses.

2.4 Does the quality of 'quantitative' research matter?

So far I have highlighted a range of conceptual issues involved in assessing the validity of experimental designs. Another feature of the literature on assessing the quality of 'quantitative' research is the body of work that has examined empirically the question of whether the design, quality of execution and analysis of studies evaluating the effects of interventions *matters* (e.g. Abraham *et al.*, 2004; Guyatt *et al.*, 2000; Moher *et al.*, 1998; Peersman *et al.*, 1999; Schulz *et al.*, 1995). Wortman (1994) credits Glass (1976) with providing the first description of how to examine the quality of studies in an empirical way. Glass (1976, p4) stated that "It is an empirical question whether relatively poorly designed studies give results significantly at variance with those of the best designed studies", and later went on to state that the "sensible course to follow is to describe – in quantitative terms – features of designs and correlate them with the study findings: the obtained relationships will reveal how important matters of design are and precisely what to do about them" (Glass, 1978, p3). Although it is not yet possible to predict the way in which the results of poorly designed studies will vary from better designed studies (Britton *et al.*, 1998; MacLehose *et al.*, 2000; Kunz and Oxman, 1998), a number of reviews have found that compared to higher quality studies, lower quality studies tend to overestimate the effects of interventions.

In a chapter entitled 'So what's so special about randomisation?' Kleijnen and colleagues (Kleijnen *et al.*, 1997), review empirical studies which have examined differences in findings between evaluations using random or non random methods for allocating participants to intervention or control groups. On average these studies find that compared to properly randomised controlled trials, studies using non-random allocation tend to exaggerate the beneficial effects of treatments. For example, in a study of 145 trials of treatment for myocardial infarction, Chalmers *et*

al. (1983) found that significant benefits of treatment were found in 25% of the non-randomised trials compared to 5% of RCTs. Kleijnen *et al.* (1997, p97) conclude that exaggerated treatment effects seem to be "primarily due to a poorer prognosis in non-randomly selected controls". Guyatt *et al.* (2000) have put forward a similar explanation to Kleijnen *et al.* (1997) for their finding that non-randomised studies can exaggerate the beneficial effects of pregnancy prevention interventions for young people. They argue that the most likely explanation for the result was that the young people who were non-randomly assigned to an intervention group were already predisposed to better outcomes regardless of the intervention (e.g. those most likely to use contraception were more willing to participate in the intervention).

Juni *et al.* (2001b) go beyond comparisons of randomised and non-randomised trials to compare high quality and low quality RCTs. They pooled the results of four methodological studies examining whether biases present in RCTs can lead to different conclusions about the effects of healthcare. Their study found that compared to trials which had adequately protected against selection bias, the results of trials with inadequate protection against selection bias – indicated by a failure to conceal the allocation process so that those allocating had some degree of control over who goes into the intervention or comparison group - showed a greater benefit of treatment. However, Juni *et al.* (2001b) found that the impact of performance bias and detection bias on trial results were less clear, and that the impact of attrition bias has not yet been adequately studied.

The finding that lower quality studies compared with high quality studies tend to overestimate intervention effectiveness appears to be as pervasive in other areas of public policy as it is within healthcare. In the area of health promotion, for example, the number of times authors conclude that an intervention has a positive effect is much larger amongst studies with low internal validity, regardless of whether the

interventions under study are implemented in the workplace to promote health; aimed at promoting the sexual health of young people; or involve interventions delivered by young people for promoting the health of other young people (Harden *et al.*, 1999a,b; Oakley *et al.*, 1995; Peersman *et al.*, 1996, 1998, 1999). For example, Peersman *et al.* (1999) compared author conclusions about effectiveness in all identified trials of workplace health promotion interventions to reduce cholesterol levels with conclusions from the sub-set of these trials which had been assessed as well-designed and reported (comparable intervention and control groups, presentation of pre and post outcome data and reporting on all outcomes). Authors concluded positive effects of interventions for reducing cholesterol levels in 70% of all trials, but this dropped to 58% for well-designed trials.

As Oakley (2000, p311) notes, the results of these studies should not be dismissed as the work of those engrossed in highly specific methodological work which has no wider relevance. She states that such work is of "great *substantive* importance" (emphasis in original) and that "The lesson that interventions are often less effective when subjected to well-designed evaluation than they seem to be without this means that decisions about policy and practice can be taken on the basis of better evidence and more realistic expectations about the differences that any particular intervention is likely to make." Thus evidence from well-deigned evaluation studies serves as a 'reality check' for the overly optimistic claims that are often made about new approaches to solving health, educational or other social problems.

2.5 Assessing the quality of 'quantitative' research in systematic reviews

Methodological work that has found systematic differences between the findings of high quality studies compared to low quality studies provides strong justification for assessing the quality of studies within a systematic review. As noted in chapter 1, a

key principle underlying the methods of a systematic review is to reduce any bias present in the primary studies it includes. Another principle is to reduce bias and error in the application of the review methods themselves. These principles raise questions about how systematic reviewers should make quality assessments and how their results should be used in a systematic review. The literature on 'quantitative' research suggests that we should use standardised tools or checklists, which assess study qualities that have been related, either empirically or theoretically, to biased effect estimates (Juni *et al.*, 1999; Moher *et al.*, 1995). The literature also suggests that we should examine the relationship between study quality and effect sizes, and downgrade or exclude those studies with features related to biased effect sizes (Juni *et al.*, 2001a,b; Moher *et al.*, 1999).

a) Tools to assess the quality of 'quantitative' research in a systematic way

There are many tools available for assessing the quality of RCTs. Moher *et al.* (1995) identified a total of 25 published up until 1993 and Juni *et al.* (1999) found a further 14 tools published up to 1997. The number of items within tools identified by Moher *et al.* (1995) ranged from three to 34. Each tool provided a scoring system or scale to provide an estimate of the extent to which trials had minimised the introduction of bias. Depending on the tool, trials could be assigned a minimum score of -10 and a maximum score of 100. All but nine of the tools specified a 'threshold' to identify trials of high quality. Between them the tools covered the four different types of validity described by the social scientists Cook and Campbell (1979) - internal validity, statistical conclusion validity, external validity and construct validity, as well as other aspects such as trial organisation. However, they varied in the weight given to different quality dimensions or to different aspects of the same quality dimension. For example, scores on items concerning adequate

randomisation and/or concealed allocation from the domain of internal validity, could contribute anything between 3% to 40% of the total quality score.

With this variation, it is perhaps not surprising that applying these tools to the same trials can give different results on their quality. This was found in a study examining 17 trials comparing two different types of a drug for use in preventing deep vein thrombosis in surgery (low molecular heparin and standard heparin) (Juni *et al.*, 1999). Across the tools, the number of trials rated as high quality ranged from three to 16, and the number rated as low quality ranged from one to 14. Juni *et al.* (1999) also found that different effect estimates were produced depending on which tool was used to include or exclude studies. The high quality trials identified by some tools showed that there was no difference between the two drugs in preventing deep vein thrombosis, whilst low quality trials showed low molecular heparin to be superior. However, with other tools the opposite was the case. Trials of high quality revealed low molecular heparin to be superior, whilst low quality trials showed that there was no difference between the two drugs.

Juni *et al.* (2001a,b) argue that there are weaknesses in many of the tools they reviewed because they include items which are not relevant to assessing internal validity, an aspect of quality which they see as being of prior importance to any of the other dimensions of validity. For example, with one of the first tools to be developed (Chalmers *et al.*, 1981), a composite quality score is calculated from scores on individual items covering dimensions of external validity (e.g. were interventions described?), trial organisation (e.g. were starting and stopping dates of the trial provided?); and presentation of data (e.g. were test statistics and p values provided?), as well as items assessing internal validity. With the summary score reflecting scores on all these dimensions, trials which have low internal validity may score just as highly as those with high internal validity as the former may have scored higher on other dimensions. Juni *et al.* (2001a) also note that for those tools

which are solely focused on internal validity, some include additional items for which there is little evidence that they play a role in the production of biased estimates of effect sizes. Furthermore some tools do not include items for which there *is* evidence they do play a role in the production of biased estimates of effect sizes. For example, the widely used tool developed by Jadad and colleagues (Jadad *et al.*, 1996), includes an item on whether allocation sequences were appropriately generated which has not been consistently been found to play a role in the production of biased estimates, but do not include anything on whether allocation was adequately concealed which has.

b) How should assessments of quality be used in a systematic review?

There are three main ways in which the quality of individual studies can be taken into account within a systematic review (Juni *et al.*, 2001a,b; Moher *et al.*, 1999). The 'threshold' approach excludes from the review studies that do not meet an acceptable level of quality. This means that only the findings of those studies judged to be of the highest quality can contribute to the conclusions of the review. In the 'weighting approach' studies are assessed and labelled according to levels of quality. However, in contrast to the threshold approach, all study findings are allowed to contribute to the review conclusions, but studies of lower quality are allowed to contribute less. The third approach involves the employment of a 'sensitivity analysis' to examine whether variation in the findings of individual studies can be accounted for by variations in methodological quality. On the basis of the results of this analysis, a decision can be made on whether to focus interpretation on those studies which show strengths on the items which have been identified as protecting against the production of biased effect estimates (Moher *et al.*, 1999).

As highlighted above, some tools enable reviewers to produce an overall rating of the quality of each trial on a scale. Trials are rated on a number of items to assess quality and these are added up to produce a summary score. This is known as the 'composite' approach to using quality assessments in a systematic review. However, when this approach is used in a sensitivity analysis interpretation of the results can be difficult. In an article entitled 'Quality scores are useless and potentially misleading', Greenland (1994) outlined three potential explanations for finding no association between quality scores and effect estimates. The first possibility is that there is no association. The second is that there is in fact an association between some aspects of trial quality and effect estimates, but this is 'drowned' in the summary score. Thirdly, it might be that again there is an association between two or more aspects of trial quality and effect estimates, but these 'cancel each other out' (i.e. one aspect may be negatively correlated, the other positively correlated). A 'component' approach is therefore recommended whereby the association of each item (or a sub-set of items) within a tool is examined separately for its association with effect estimates (Juni *et al.*, 2001a,b; Wortman, 1994). Juni *et al.* (2001a, p100) argue that such an approach "takes into account that the importance of individual quality domains and the direction of potential biases associated with these domains, will vary between the contexts in which trials are performed".

This point is well illustrated by the study described previously of 17 trials comparing two different types of a drug for use in preventing deep vein thrombosis in surgery (Juni *et al.*, 1999). Using a component approach, this study found that blinding of outcome assessors was the only aspect of quality significantly associated with effect estimates - when trials did not use blinding of outcome assessors, treatment effects of low molecular weight heparin were exaggerated by 44%. Thus, in this particular context (drugs for preventing deep vein thrombosis in surgery), adequate

concealment of allocation and the use of intention to treat analysis did not play a role in the production of biased effect estimates. In an explanation which shows remarkable similarity to the idea from the Campbellian framework of specifying and then ruling out 'plausible threats to validity', Juni *et al.* (2001a, p102) comment "The importance of blinding outcome assessors could have been anticipated because the interpretation of fibrinogen leg scanning, the test used to detect deep vein thrombosis, can be subjective" and that the importance of allocation concealment "may to some extent depend on whether strong beliefs exist among investigators regarding the benefits or risks of assigned treatments.....strong beliefs are probably more common in trials comparing an intervention to placebo than in trials comparing two similar, active interventions".

2.6 Summary and conclusion

In this chapter I have presented and discussed some of the major themes in the literature on the topic of quality in 'quantitative' research addressing questions about the effects of interventions. Two important frameworks in this literature for thinking about quality in research are 'threats to validity' and 'sources of bias'. Central to each framework is the assessment of quality through attention to the ways in which study design and methods may lead to erroneous answers to study questions. In the 'threats to validity' framework, studies evaluating the effects of interventions are assessed according to whether faults in study design and study methods provide a more plausible explanation for any observed effects rather than the intervention under test. In the 'sources of bias' framework, studies are quality assessed according to the extent to which study design and methods have been able to minimise the introduction of bias into effect sizes.

Another important feature of the literature is the work that has tested whether the quality of 'quantitative' research impacts on its findings. Knowledge about how to assess the quality of experimental research in systematic reviews has advanced through this empirical work because such work has helped to distinguish between significant (e.g. selection bias) and non-significant (e.g. trial organisation) errors for answering questions about effectiveness. These findings have been invaluable for guidance on the design and conduct of trials and for tools to assess the quality of trials. Work examining whether and how study quality impacts on study findings is notably absent in the literature relating to quality in 'qualitative' research. This issue is discussed in the next chapter.

CHAPTER 3

A review of the conceptual and methodological literature relating to the topic of quality in 'qualitative' research

3.1 Aims and rationale

In the review reported in this chapter, I aim to identify and discuss some of the major themes in the literature on the topic of quality in 'qualitative' research. Like the review reported in chapter 2, at first I focus on the quality of 'qualitative' research in general and then move on to consider quality assessment with particular reference to systematic reviews. To identify relevant literature, I undertook searches on six bibliographic databases and four specialist registers by combining terms for 'qualitative' research (e.g. 'qualitative', 'ethnography', 'in-depth interviews') with terms for 'quality' (e.g. 'quality', 'validity', 'standards')[1]. I undertook my literature search in 2002, a time when there was very little literature on assessing the quality of 'qualitative' research in systematic reviews. I therefore supplemented my original search with a search for this specific literature at regular intervals from 2002 to 2006.

As might be expected on a topic that has stimulated much debate and controversy, there is a large literature and I did not attempt to review every single paper I identified. My aim was to compare and contrast the range of perspectives that exist This means that I do not refer to authors that simply summarise the perspectives of other authors for audiences in a different field (e.g. Ackroyd, 1996; Gilner, 1994; Sparkes, 1998; Sykes, 1990). Because I searched databases indexing literature

[1] The details of this search are reported in full in chapter 5 as the search was the same one that I used to identify the quality assessment tools in my first new methodological study.

from the social and health sciences, I identified literature spanning a number of different disciplines including education, nursing, psychology and sociology. I did not, however, search explicitly within the anthropological literature and so I cannot claim to have represented all perspectives on quality in 'qualitative' research (e.g. Geertz, 1983, Leach, 1982). On the other hand, some of the literature I identified did draw on anthropological work so it us unlikely that I have missed perspectives from anthropology altogether (e.g. Kirk and Miller, 1984; LeCompte and Goetz, 1982).

Within the literature on quality in 'qualitative' research there are debates about whether the quality of 'qualitative' research should be assessed, and if so how. In the realm of systematic reviews, there are additional debates about when quality assessment should take place in the systematic review process, whether studies should be excluded from reviews on the basis of quality, and whether quality should be assessed using a 'tool' or 'checklist' or by expert judgement alone. All of these debates will be discussed, but the chapter begins by outlining three positions on whether and how the quality of 'qualitative' research should be assessed.

3.2 Three positions on quality in 'qualitative' research

Not everyone agrees that the quality of 'qualitative' research can or should be assessed. Like other reviews of the literature on the quality of 'qualitative' research, I identified several positions on whether and how the quality of 'qualitative' research should be assessed (Angen, 2000; Devers, 1999; Hammersley, 1992; Madill *et al.*, 2000; Murphy *et al.*, 1998; Oakley, 2000; Spencer *et al.*, 2003). The first position, nearly always labelled by others as the 'conventional' position, is that 'qualitative' research should be judged according to the same standards as 'quantitative' research. The second, often named as the 'alternative' position, is that 'qualitative' research should be assessed using completely different standards to 'quantitative'

research. The third position, named by Madill *et al.* (2000) as a 'radical' position, is that 'qualitative' research should not be judged by fixed standards, especially those that judge quality through method.

a) The 'conventional' position

Kirk and Miller (1986) and LeCompte and Goetz (1982) are two classic expressions of this position. These authors argue that any type of scientific research, whether 'qualitative' or 'quantitative', should be judged according to the reliability and validity of its findings. Both sets of authors see validity as concerned with the accuracy of scientific findings and reliability as concerned with the replicability of scientific findings. Kirk and Miller (1986, p24) see validity as "calling things by their right names". LeCompte and Goetz (1982, p32) define validity as the extent to which "constructs devised by researchers represent or measure the categories of human experience that occur". Both pairs of authors argue that whilst 'qualitative' research can show strengths in terms of validity, reliability - which requires study methods to be reported in enough detail for replication - is a major problem because of: a lack of common descriptors for the techniques used to collect data; "vague, intuitive and personalistic" analytical processes; and a traditional focus on "artful presentation" of results rather than detailed description of methods (LeCompte and Goetz, 1982, p36). On the former point, Kirk and Miller (1986) argue that the fieldwork stage in ethnographic research - characterised by prolonged engagement with the people under study and continual testing and revision of emerging hypotheses – functions as a validity check because it has an 'in-built' sensitivity to any discrepancies between the concepts of the researchers and the lives and understandings of the participants.

In the 'conventional' position, the reliability and validity of 'qualitative' research findings are judged according to how well the study was carried out and reported. Kirk and Miller (1986, p73) argue that the "the problem of validity" should be handled through fieldwork and the "problem of reliability" should be handled by "documented ethnographic decision-making". As noted above, Kirk and Miller (1986) see fieldwork as a check on validity. To handle the "problem of reliability", Kirk and Miller (1986) recommend that 'qualitative' researchers document their decision-making according to four phases that are applicable to any type of scientific activity: invention (preparation and research design to produce a plan of action); discovery (data collection to produce information); interpretation (data analysis to produce understanding); and explanation (communicating the findings of the research).

LeCompte and Goetz (1982) are influenced explicitly by Cook and Campbell's 'threats to validity' approach. They describe a whole range of threats to reliability and validity for 'qualitative' researchers to be aware of and recommend strategies for guarding against such threats. Threats include those arising from: the social position of the researcher in relation to the group being studied; the personal and disciplinary biases of the researcher; who the informants are and why they were chosen; the social situations in which data were generated; and the theoretical constructs and assumptions that informed the research. Strategies for reducing these threats include: the provision of many 'low-inference descriptors'[2]; the use of multiple observers to establish consensus on what has been observed; using participant informants to check the observations of researchers; subjecting the observations of the researchers to peer review; and the use of 'mechanical devices' to record and preserve raw data.

---

[2] 'Low inference descriptors' are examples of primary data which might be quotes from participants or extracts from field notes.

b) The 'alternative' position

In the 'alternative' position it is argued that 'qualitative' research should not be assessed by the same criteria as 'quantitative' research. Educational researchers Egon Guba and Yvonna Lincoln are two influential exponents of this position (e.g. Guba and Lincoln, 1981; Lincoln and Guba, 1985). Guba and Lincoln view 'quantitative' research, and the associated standards of reliability and validity, to be based on an outdated 'positivist' model of science. Such a model, they argue, naively assumes that there is a single, tangible, unchanging reality that exists independently from the research that aims to represent or approximate it. In contrast, the 'qualitative' paradigm assumes multiple versions of reality that shift over time. Rather than seeking accuracy in the findings of 'qualitative' research or the best approximate representation of a single 'truth', Guba and Lincoln see the task of research as representing multiple constructions of reality. Ultimately, Lincoln and Guba (1985) see 'qualitative' and 'quantitative' research as competing paradigms.

Lincoln and Guba (1985) take the four standards that they consider to be characteristic of the 'positivist' or 'quantitative' paradigm – 'internal validity', 'external validity', 'reliability' and 'objectivity' – and offer a set of alternative standards:

1) *Credibility*: the inquiry must show that multiple constructions of reality are represented adequately;
2) *Transferability*: the inquiry must offer working hypotheses about the phenomenon under study with a detailed description of the context and time in which they were found to hold;
3) *Dependability*: the inquiry must establish the acceptability of the processes of the inquiry used to collect and analyse data; and

4) *Confirmability*: the inquiry must show that the interpretations are supported by and grounded in the data rather than the researcher's personal constructions.

Lincoln and Guba (1985) propose various techniques for enhancing the credibility, transferability, dependability, and confirmability of 'qualitative' research. They argue that dependability and confirmability can be established by the provision of an audit trail, a detailed account of what happened at all stages of the research. 'Thick description'[3] must be provided to enhance an assessment of transferability, and prolonged engagement with the phenomenon under study, triangulation of different sources of data, negative case analysis (revising a hypothesis until it accounts for all cases), and member checking (asking study participants to check findings) are all examples of the techniques that Lincoln and Guba (1985) propose for establishing credibility. Lincoln and Guba (1985, p329) were keen for others to try out their quality assessment proposals and believed that the way forward would be an "empirical matter":

> "there is still a major gulf between the theoretical definitions of the trustworthiness criteria and the means of operationalising them. It is one thing to suggests that triangulation is needed, for example, and quite something else to say how much, or what type of triangulation will suffice to establish a minimally acceptable level of trustworthiness . . .It seems likely that the development of operational means and decision rules for these various criteria and the techniques related to them will be an empirical matter; only through efforts to apply the criteria will the field come to an understanding of what decision rules make sense."

[3] What constitutes 'thick description' is not well defined in Lincoln and Guba (1985). They suggest that 'thick description should comprise of "a thorough description of the context or setting within which the inquiry took place and with which the inquiry was concerned" and "a thorough description of the transactions or processes observed in that context that are relevant to the problem, evaluand, or policy option" (Lincoln and Guba, 1985, p362).

The work of Lincoln and Guba has been hugely influential in the debate about how the quality of 'qualitative' research should be assessed. Earnest appeals have been made against the application of 'quantitative' criteria to 'qualitative' research (e.g. Altheide and Johnson, 1998; Leininger, 1994; Yoneg and Stwein, 1988); credibility, transferability, dependability and confirmability are used to structure discussions about quality in 'qualitative' research (e.g. Beck, 1993; Elder and Miller, 1995) and the use of audit trails, thick description, triangulation and member checking are routinely cited as techniques for enhancing the rigour of 'qualitative' research (e.g. Creswell and Miller, 2000; Giancomini and Cook, 2000a,b; Long and Godfrey, 2003; Miles and Huberman, 1994; Treloar, 2000; Whittmore *et al.*, 2001;). However, despite the intention to articulate 'alternative' quality criteria, the techniques suggested by Lincoln and Guba (1985) for enhancing credibility, transferability, dependability and confirmability show remarkable overlap with the procedures for establishing reliability and validity suggested by those from the 'conventional' position. This problem of just how 'alternative' the criteria proposed by Lincoln and Guba are has been noted by those who represent the 'radical' position in the debate about assessing the quality of 'qualitative' research.

c) The 'radical' position

In the 'radical' position it is argued that it is not possible to set non-arbitrary criteria or standards to assess the quality of qualitative research. A key exponent of this position is John K Smith (e.g. Smith, 1984; Smith, 1993). Smith argues that the 'alternative' criteria that Lincoln and Guba (1985) propose represent 'foundational' standards that are inconsistent with the 'anti-foundational' assumptions of multiple realities and 'truth' as a "socially and historically conditioned agreement" (Smith, 1984, p380). Smith adopts a relativist or anti-realist position and denies the possibility that we can judge research according to how accurately it reflects reality.

Smith argues that social inquiry should be based on a naturalistic (or 'qualitative') paradigm which assumes that reality is "mind-dependent and that there are multiple realities" (Smith, 1984, p386). Within such a paradigm there are "no procedures or criteria exclusive to or particularly appropriate for social inquiry" (Smith, 1984, p390). From this perspective all research findings are simply interpretations rather than claims about "how things really are" (Smith, 1984, p390).

Another key feature of Smith's position is the rejection of methods and procedures as a route to reliable knowledge. Citing Rorty (1982), Smith argues that we should "dispense with the traditional ideas of objectivity and truth and realise that we are 'beyond method'" (Smith, 1984, p390). Although Smith argues against the application of foundational methodological standards separating out trustworthy from untrustworthy results, he does still suggest that there are ways to distinguish between good and bad research:

> "In the end, the task of distinguishing knowledge from opinion and good from bad research is an eminently practical and moral task – not an epistemological one whose rationality is directed by more or less determinate rules or standards" (Smith, 1993, p163).

Smith's call to go 'beyond method' to distinguish between good and bad research has been taken up by others who adhere to an anti-realist ontology and a relativist epistemology[4] (e.g. Lather, 1993; Lincoln, 1995; Schwandt, 1998). For example, Lincoln (1995) - who now rejects her earlier 'foundational' position in her work on

---

[4] Ontology refers to the nature of the world and epistemology refers to the nature of knowledge. Those adopting an anti-realist ontology take the view that 'things' or 'entities' in the world are always socially constructed - they do not have a 'real' existence that is independent of the way we think about them. Those adopting a relativist epistemology take the view that all knowledge claims are relative to each other or to a particular context and cannot be judged as 'true' or 'false' by appeal to whether or not they accurately represent an external reality.

'alternative' criteria with Guba - argues that 'qualitative' research should be judged against 'emerging' standards that are said to "recognise and validate relationships between the inquirer and those who participate in the inquiry" (p278). These emerging standards - or 'ethical validation' (Angen, 2000) - suggest that 'qualitative' research needs to be judged in terms of its ability to: bring about action and promote social justice; promote an equitable context in which all voices are heard; and develop solutions to practical problems. In contrast to the 'conventional' and 'alternative' positions, the 'radical' position appears to emphasise an assessment of the quality of the findings of 'qualitative' research (e.g. do the findings provide a solution to a practical problem?) rather than an assessment of how well the 'qualitative' study has been carried out.

3.3 The 'checklist' debate

The literature on quality in 'qualitative' research has seen an increasing number of 'checklists' designed to help journal editors, peer reviewers and readers critically appraise 'qualitative' research. I subject these checklists to in-depth analysis later on in this thesis (see chapter 5), but in this section I briefly discuss three major concerns that have been raised in the literature about their use. The first concern is whether the technical procedures advocated by some checklists (e.g. triangulation and respondent validation) are actually misguided as strategies for enhancing validity in 'qualitative' research (Barbour, 2001; Bloor, 1997; Oakley, 2000). (As Murphy *et al.*, (1998) note, these strategies are better considered as techniques for collecting a wide range of data and perspectives.) The second concern surrounds the emphasis in checklists on the procedural aspects of research (how it is done) at the expense of any consideration of the quality of the insights offered by the findings (Eakin and Mykhalovskiy, 2003). As Barbour and Barbour (2003) note, although

systematic and rigorous methods should lead to valuable theoretical and conceptual insights, this is not always the case.

The third and final concern has been discussed by Barbour (2001, p1115) who argues that checklists may reduce 'qualitative' research to a list of "technical procedures" such as triangulation, respondent validation, and purposive sampling. The fear here is that the conduct and evaluation of 'qualitative' research will be driven by uncritical and prescriptive application of a checklist rather than on a broader understanding of 'qualitative' research design and analysis. Implicit here is the suggestion that judging the quality of 'qualitative' research is a complex task that is best done by experienced 'qualitative' researchers. Sandelowski *et al.* (1997, p369) explicitly subscribe to this view and argue that those judging 'qualitative' research have to be "true 'connoisseurs'…of qualitative research to distinguish between surface errors and mistakes fatal enough to discount findings". The debate about checklists reminds us that assessing the quality of 'qualitative' research is not only a contested area because of different views on the most appropriate quality criteria. It is also a contested area because of different views about who is qualified to make judgements about the quality of 'qualitative' research.

3.4 Is the issue of quality really so different for 'qualitative' research?

So far I have outlined three positions on assessing quality in 'qualitative' research and described the concerns of 'qualitative' researchers regarding the use of 'checklists' to assess quality. The 'conventional' position on assessing quality in 'qualitative' research suggests that the issue of judging quality is the same regardless of whether the research is 'qualitative' or 'quantitative'. The 'alternative' position suggests that it is inappropriate to judge 'qualitative' research by 'quantitative' standards because 'qualitative' research is based on different

ontological and epistemological assumptions. The 'radical' position suggests that the 'alternative' position does not go far enough away from the 'conventional' position. Those adopting a 'radical' position argue that it is inappropriate to judge 'qualitative' research according to any kind of methodological standard because this kind of research is based on an anti-realist ontology and a relativist epistemology. The argument from the 'radical' position that it is not possible to set quality assessment criteria for 'qualitative' research is difficult to sustain for several reasons. Firstly, it falls down in the light of the fact that not all 'qualitative' research is underpinned by relativist assumptions (Murphy *et al.*, 1998). Secondly, as Bailey (2001) reminds us, the 'radical' position is self-contradictory; truth is rejected at the same time as making a truth claim about how it is impossible to set criteria to assess quality. Thirdly, the 'radical' position presents research and academia as forums for "for political whim and fancy" (Bailey, 2001, p170) by putting political goals above truth-seeking goals (Bailey, 2001; Hammersley, 2000a; Murphy *et al.*, 1998). Similarly, the 'alternative' position is difficult to sustain if one does not accept the competing paradigm view of 'qualitative' and 'quantitative' research. Those advocating that it is not appropriate to assess the quality of 'qualitative' research with 'quantitative' criteria often rely on an inaccurate characterisation of the model of science that 'quantitative' research is based upon. According to advocates of the 'alternative' position (e.g. Lincoln and Guba, 1985), 'quantitative' research is based on a 'naïve realist' or 'positivist' model of science which assumes that we can perceive the world exactly as it exists (Mark, 2000). However, not all 'quantitative' research is based on such a model of science. For example, Donald Campbell who developed the framework for assessing the validity of experimental research discussed in chapter two, was a critical (rather than naïve) realist who assumed that our knowledge of the world would always be imperfect (Mark, 2000). As already noted, a further problem with the 'alternative' position is that despite its claim to be different, its content shows remarkable similarity with the 'conventional' position.

If one accepts that the scientific method – a method which assumes the fallibility of knowledge and truth as a regulatory ideal - is appropriate for both 'qualitative' and 'quantitative' research, it makes sense to suggest that all research should be judged in the same basic way (Hammersley, 1992; Murphy *et al.*, 1998; Oakley, 2000). Hammersley (1992), who views the function of research as providing "information that is both true and relevant to some legitimate public concern" (p68), argues that quality should be assessed according to validity and relevance. For Hammersley (1992, p67), assessing validity involves making judgements about "the likelihood of error" (p67). Similarly, Oakley (2000, p72) suggests that 'qualitative' and 'quantitative' research should both be judged according to some common standard and that "one might reasonably argue that the distinguishing mark of all 'good' research is the awareness and acknowledgement of error". In this light, the 'conventional' and 'alternative' positions, and the 'checklists' which have been compiled to aid assessments of quality, offer a starting point for thinking through all the potential sources of bias and error in 'qualitative' research.

3.5 Assessing the quality of 'qualitative' research in systematic reviews

The debate in the social science literature about how the quality of qualitative research should be assessed has presented problems for those trying to include 'qualitative' research in systematic reviews. In systematic reviews of the effects of healthcare carried out by the Cochrane Collaboration study quality is a key criterion for study inclusion. Although more detailed quality assessments may take place later, studies are usually included or excluded from a Cochrane review on the basis of the presence or absence of what is considered to be a fatal flaw in studies examining the effects of interventions: adequate randomisation of participants into intervention and control groups. It has been argued that the lack of consensus in the

literature on assessing the quality of 'qualitative' research makes it impossible to define equivalent criteria for selecting high quality qualitative studies to include in reviews (Daly *et al.*, 2006; Dixon-Woods *et al.*, 2006; Popay, 2005).

Despite these difficulties there is broad agreement amongst those working on the problem of how to review 'qualitative' research in a systematic way that quality should be assessed, even amongst those who propose methods for systematically reviewing 'qualitative' research in its own right (e.g. Paterson *et al.*, 2001; Noblit and Hare, 1988; Sandelowski and Barroso, 2007). For example, Noblit and Hare (1988) recommend that ethnographic studies are quality assessed according to the adequacy of the metaphors used to communicate study findings. However, there is debate about a) whether to include or exclude studies on the basis of quality; a) how to assess quality; c) whether quality should be assessed prior to or during the synthesis stage of the review; and d) whether the 'quantitative' model of reviews will fit 'qualitative' research.

a) Should studies be included or excluded on the basis of quality?

On this first issue, Sandelowski *et al.* (1997) argue that studies should not be excluded on the basis of quality because of the lack of agreement on what constitutes a high quality or low quality 'qualitative' study. Others, however, argue that poor quality 'qualitative' studies should be excluded to avoid distorting the review results (Dixon-Woods *et al.*, 2004a,b) and to ensure that review users are able to draw on reliable evidence (Attree *et al.*, 2006). Dixon-Woods *et al.* (2004a,b) also argue that low quality studies should be excluded in order to avoid undermining the credibility of reviews that include 'qualitative' research.

b) How should the quality of 'qualitative' research be assessed in reviews?

Amongst those systematic reviews that have included 'qualitative' research, a common solution to the problem of how to assess quality has been to use a standardised checklist or tool[5]. For example, Campbell *et al.* (2003) used a critical appraisal tool in a review of studies on patient experience of diabetes and diabetes care, as did Attree (2004) in a review of studies on child poverty, and Milton and Whitehead (forthcoming) in a review of studies on the social consequences of children's ill health. These authors report several advantages of using a checklist to help them judge the quality of studies identified as relevant for their reviews. Attree and Milton (2006) noted that the use of a checklist in their reviews provided a thorough and systematic basis for comparing the strengths and weaknesses of different studies and stimulated debate amongst reviewers. Campbell *et al.* (2003) found that in addition to helping weed out inappropriate and poor quality papers, the use of a checklist acted as a first stage for synthesis by helping reviewers to engage with studies and identify key concepts from study findings.

Whilst some believe that checklists can be useful if used in a critical and flexible way, others reject checklists altogether. Some reject checklists because they view them as dangerous "cluster bombs" from the "arsenal of the quantitative camp" (Jones, 2004; p95). Others reject checklists because they take the view that the meaning and quality of research will only emerge in the interaction between the findings and the critical reader (Garrett and Hodkinson, 1998). From this perspective, any attempt to apply predetermined and fixed criteria is illogical.

c) Can quality be assessed prior to the synthesis stage of a review?

---

[5] In systematic reviews of trials it is standard practice for reviewers to evaluate the quality of studies using a tool or a checklist to prompt judgements on whether or not studies have taken steps to minimise the introduction of bias and error. This helps to ensure that reviewers treat each trial in the same way.

Those who argue that studies should be excluded from reviews on the basis of quality assume that it is possible that quality can be assessed prior to the synthesis stage of a review. Others, however, have challenged this view. Noblit and Hare (1988) – who developed a method for synthesising the findings of 'qualitative' research called meta-ethnography – adopt a similar view to Garrett and Hodkinson (1998) and argue that the quality of 'qualitative' research will only emerge in the synthesis stage of a review. According to Noblit and Hare (1988), and more recently, Pawson (2006b), judging research quality is not just about determining whether the research has been carried out according to sound procedures. Judging research quality is also about examining how study findings fit (or do not fit) with the findings of other studies. How study findings fit with the findings of other studies cannot be assessed until the synthesis is completed. As Pawson (2006b, p141) puts it, the "worth of a study is determined in the synthesis".

d) Will the 'quantitative' model of reviews fit 'qualitative' research?

Underlying all the debates discussed so far is a broader debate about whether a 'quantitative' model of systematic review will fit 'qualitative' research. Those who have used checklists to include or exclude studies prior to synthesis represent the view that the 'quantitative' model of systematic review can be applied to 'qualitative' research. Others have been highly critical about whether the 'quantitative' approach, should be applied to 'qualitative' research (Barbour and Barbour, 2003; Booth, 2001; Dixon-Woods *et al.*, 2006; Jones, 2004; Noblit and Hare, 1988). These authors argue that it is inappropriate to try to make a 'quantitative' template for doing systematic reviews 'fit' and that a distinctive 'qualitative' approach is needed[6]. Such a distinctive approach would involve questions specified in broad terms, to act

[6] The 'quantitative' template for doing systematic reviews is typically assumed to be the one described in the *Reviewers Handbook* (Higgins and Green, 2006) produced by the Cochrane Collaboration.

as 'compasses', as opposed to questions specified in narrow terms to act as 'anchors'; purposive sampling rather than exhaustive searching; quality appraisal as part of the synthesis rather than as a pre-curser; the use of unprompted expert judgement rather than a standardised checklist; and synthesis as 'interpretations' rather than 'aggregations'.

There are echoes of the social science paradigm wars in the characterisation of the systematic review as a 'quantitative' approach and the development of a distinctive 'qualitative' approach to the review of 'qualitative' research. Indeed, the use of 'quantitative' and 'qualitative' to describe approaches to synthesis is beginning to permeate the literature on research synthesis, with 'quantitative' approaches usually cast as the 'villain' and 'qualitative' approaches as the 'hero' (e.g. Booth, 2001; Dixon-Woods *et al.*, 2006; Jones, 2004). For example, Jones (2004, p98) argues that we should abandon the "tyranny of numbers" to embrace a more inclusive 'qualitative' approach. Re-creating the social science paradigm wars in research synthesis may have the same negative consequences that have been identified within primary research (described in chapter 1). As Darbyshire (1997) has noted in relation to primary research, 'quantitative' approaches to research synthesis may be dismissed as inappropriate for the study of the social world, whilst 'qualitative' approaches to research synthesis become the new orthodoxy. A paradigm approach may also: overlook the similarities between 'qualitative' and 'quantitative' approaches to research synthesis; exaggerate the differences; and overlook the possibility that 'quantitative' and 'qualitative' approaches can be complementary, rather than competing, and that selection of one approach over the other could depend on the type of review question asked.

3.6 Summary and conclusion

In this chapter I have discussed and evaluated three divergent positions in the literature on quality in 'qualitative' research, described the concerns of 'qualitative' researchers regarding the use of 'checklists' to assess quality, and raised the debates about assessing the quality of 'qualitative' research in the specific literature on 'qualitative' research in systematic reviews. My review suggests that the literature in this area – both in general and more specifically in connection with systematic reviews - has been dominated by debates about whether 'quantitative' approaches to assessing quality can be applied to 'qualitative' research. Moreover, there is evidence to suggest that the paradigm wars are being re-created within the systematic review literature. The paradigm wars position 'qualitative' approaches to systematic review against 'quantitative' approaches, and suggest that the 'quantitative' approach to systematic reviews cannot be applied to 'qualitative' research.

I have questioned whether the issue of quality in 'qualitative' research is really so different from the issue of quality in 'quantitative' research and want to raise a similar question in relation to whether the principles of reviewing 'qualitative' research in a systematic way should be different to reviewing 'quantitative' research in a systematic way. I believe that it is worth trying to apply what some have called a 'quantitative' model of systematic reviews to 'qualitative' research. To avoid recreating the paradigm wars, however, I want to reframe the debate about whether or not the 'quantitative' model of systematic reviews fits 'qualitative' research, to a debate about whether or not the systematic review model developed to answer questions of effectiveness fits other questions about intervention processes and studies of people's perspectives and experiences. I do not expect there to be a perfect fit between the 'quantitative' model and 'qualitative' research. I consider the key challenge to be to find out what aspects of the model do fit and which aspects do not.

Another important feature of the literature reviewed in this chapter is the absence of any empirical work which examines how the different approaches proposed for assessing quality work when they are actually applied to 'qualitative' research. It appears that the recommendations made over 20 years ago by Lincoln and Guba (1985, p329) to bridge the "gulf between theoretical definitions of... trustworthiness criteria and the means of operationalising them" through empirical work have largely been ignored. To an extent, the fact that there is limited empirical work testing approaches to assessing the quality of 'qualitative' research leaves a blank canvas for the new methodological work undertaken in for the rest of this thesis. I do, however, use the debates described in this chapter to frame the three new methodological studies that I conducted. I have also drawn on the work described in chapter 2 to inform the design of these three new studies. For example, the first new study - a survey and evaluation of tools to assess the quality of 'qualitative' research - is framed by the debates in the literature on how the quality of 'qualitative' research should be assessed. Its design is also informed by studies that have identified and compared tools for assessing the quality of trials. As noted in chapter 2, such studies have helped to sort relevant from irrelevant criteria for assessing the quality of trials. The aims, design and methods of all three studies are described in the next chapter.

CHAPTER 4

Three methodological studies: aims, design and methods

The aim of this chapter is to give an overview and introduction to the aims, design and methods of the three new methodological studies that, together with the review reported in chapter 2 and chapter 3, make up the main body of work for my thesis. A detailed description of the aims, design and methods of the studies is given in the individual chapters reporting on each study: chapter 5 reports study one; chapter 6 reports study two, and chapter 7 reports study three. This chapter also describes the programmes of work in health promotion and public health (HP&PH) and education at the EPPI-Centre from which this thesis originates. Despite its origins in EPPI-Centre programmes of work and the fact that it draws on data generated by these programmes, I was solely responsible for the design, analysis and content of the work described in the thesis. Another aim of the chapter is to show how the work of the thesis and that of EPPI-Centre programmes are distinct.

4.1 Study one

Study one was designed to address the second specific aim of this thesis: to identify, compare, and evaluate the quality criteria that have been proposed to assess the quality of 'qualitative' research. I began the study by searching systematically for any type of literature related to the topic of assessing the quality of 'qualitative' research. I identified a large number of citations and, as I began to obtain the full reports, I soon realised that a systematic examination of all reports would lack both coherence and feasibility. The reports were a mixture of: critiques of the application of 'quantitative' quality concepts to 'qualitative' research and

proposals for alternative quality concepts (e.g. Lincoln and Guba, 1985; Lincoln, 1995); evaluations of specific strategies proposed to ensure rigour in 'qualitative' research (e.g. Armstrong *et al.*, 1997; Bloor, 1997); 'checklists' or 'tools' for critical appraisal (e.g. Cesario *et al.*, 2002; Mays and Pope, 1995); and general reviews of the field (e.g. Madill *et al.*, 2000; Murphy *et al.*, 1998). I chose the literature reporting tools as a focussed sample in which to compare and contrast the range of quality criteria that have been proposed to assess the quality of 'qualitative' research. In contrast to the more abstract and theoretical discussions about quality, tool authors generally offered clear expositions of the quality criteria they had proposed for assessing the quality of 'qualitative' research. The final design of study one was a survey and evaluation of tools for assessing the quality of 'qualitative' research.

For the survey, I designed a standardised form to collect data from each tool covering, for example, tool structure and content. Although I was able to use frequencies and counts to describe some aspects of the tools (e.g. number of items across tools), many of the data I collected were textual in nature. I analysed this data using two main strategies: content and thematic analysis. For example, I used content analysis to capture the reasons why tools had been developed, and thematic analysis to describe tool content. I conducted a separate exercise to evaluate the tools more specifically according to their strengths and limitations from a systematic review perspective. I convened a meeting with a group of experienced systematic reviewers to generate a list of desirable features for quality assessment tools and assessed each of the tools I had identified against this list.

4.2 Study two

Study two, which analyses the development of a new tool for assessing the quality of 'qualitative' studies in systematic reviews, was originally designed to bridge the

gap between study one and study three. When I began thinking about study two I had already completed some work for study one, and I had just begun to explore how to study the relationship between study quality and review results for study three. Study two originated in a need to describe the tool that was used to quality assess the studies under examination in study three. This tool was not one of the tools surveyed in study one, but a new tool that I had developed with colleagues at the EPPI-Centre. Because this tool was so different to the other tools surveyed in study one, it became clear that simply describing the tool would not be enough. I wanted to identify and explore the factors that had led to the development of such a unique approach in the literature. The fact that there had been little reflection on the development of other tools offered an additional rationale for study two.

The final design of study two was a retrospective analysis to identify the factors influencing the development of the tool. Relevant documentation on the tool was collected including draft versions of the tool, research proposals, e-mails amongst the team, and correspondence with funders. I treated all of this documentation as data, read each document in detail, and took notes to build up an account of the methodological development. As this account developed I began to identify the key factors that led to the development of the new tool. I then refined this list of key factors after discussion with two other members of the team who developed the tool. The analysis was driven by two main questions: i) how did the methodological development happen?; and ii) what are the lessons to be learnt for fostering methodological development in the future?

4.3 Study three

The third methodological study attempted to address the third specific aim of this thesis: to assess the relationship between the quality of 'qualitative' studies and the

findings of systematic reviews that include them. This felt like unknown territory and my initial approach to designing the study was little more than getting stuck into the reviews and having a look at what was going on. I chose six completed systematic reviews that had included 'qualitative' studies to analyse. I had conducted all of these reviews with colleagues at the EPPI-Centre as part of the HP&PH programme of work. I chose reviews in HP&PH rather than reviews in education because, although some of the education reviews included 'qualitative' studies, they were not suitable for my purposes. The authors of the education reviews had used 'qualitative' studies to address questions about intervention impact. I wanted my analysis to assess the relationship between the quality of 'qualitative' studies and the findings of syntheses of intervention processes and people's perspectives and experiences.

The final design of study three was a series of three retrospective analyses exploring whether the quality of 'qualitative' studies affected the findings of the reviews they were included in. By the time I was conducting the work for study three there were actually seven HP&PH reviews available that included 'qualitative' studies. I only used six, however, because the seventh review – on HIV-health promotion for men who have sex with men - had excluded low quality studies. In the other six reviews there was opportunity to explore the different roles high quality and low quality studies played because they had included all studies regardless of quality. In the first analysis, which focused on a review of peer-delivered health promotion, I examined whether low quality process evaluations produced different findings about the appropriateness of peer-delivered health promotion for young people. In the second and third analyses, which focussed on five reviews about the barriers to, and facilitators of, the health and health behaviours, I compared the contribution of low and high quality 'qualitative' studies within syntheses to find out about the perspectives and experiences of children and young people.

4.4 EPPI-Centre programme of work in HP&PH

a) Background

The EPPI-Centre has received funding from the English Department of Health (DH) for a programme of systematic review work in HP&PH since 1995. Since 1998 the DH has funded the EPPI-Centre for programmes of work in three year cycles. My thesis began during the first of these three year funding periods from 1998 to 2001. The programme for this funding period (and the next from 2001 to 2004) was entitled 'Field co-ordination in health promotion linked to the Cochrane Collaboration'. A large proportion of this programme was dedicated to undertaking a series of policy-relevant systematic reviews in HP&PH. Other work included co-direction of the Cochrane Health Promotion and Public Health Field and maintaining bibliographic registers of HP&PH evidence.

The start of the 1998 to 2001 programme of work coincided with significant shifts in UK government health policy. A focus on tackling inequalities in health and a call for individual, community and government action to improve health were strong themes of the 1999 UK government white paper on health *Saving Lives* (Department of Health, 1999a). 'Social exclusion' became an important policy concept for the UK government, defined as "a shorthand term for what can happen when people or areas suffer from a combination of linked problems such as unemployment, poor skills, low incomes, poor housing, high crime, bad health and family breakdown" (Social Exclusion Unit, 2001, p11). Improving the health of 'socially excluded groups' (e.g. the unemployed; those on low-incomes; those in care; and some ethnic minority communities) was promoted as a key strategy for reducing inequalities. These themes were reflected in the proposals for EPPI-Centre work.

For, example, suggested products for the 1998 to 2001 programme were reviews with a specific focus on socially excluded groups.

Another strong theme in *Saving Lives* was a commitment to evidence-based policy and practice. Alongside *Saving Lives* the Government launched a ten year modernisation plan for the health service in England. The consultation document, *A First Class Service: Quality in the new NHS*, made it clear that evidence-based practice would be a cornerstone of modernisation, with a vision that clinical decisions "should be based on the best possible evidence of effectiveness" (Department of Health, 1998, p2). The National Institute for Clinical Excellence (NICE) was set up in 1999 as a special health authority to develop national guidance on treatment and care based on the best available evidence. This remit was extended in 2005 to cover the promotion of good health and the prevention of ill health and NICE became the National Institute for Health and Clinical Excellence.

Within its commitment to evidence-informed health policy and practice, the UK Government also recognised the challenges in developing and collating the evidence-base, particuarly in relation to public health. The research and development strategy for public health published in 2001 outlined the problem thus:

> Knowledge to improve health and well-being .....derives from a very wide range of methodologies and approaches ..... laboratory experiments, techniques of environmental measurement and assessment, epidemiological approaches, surveys, controlled intervention studies, clinical trials and a range of other quantitative and qualitative observational and experimental methods. The challenge is to develop and extend the evidence base and to increase its usefulness (Department of Health, 2001a, p11).

Another cornerstone of the ten year modernisation plan for the health service and the public health research strategy was a commitment to listening to, and involving, patients and the public. A new national survey of patient and user experience was announced to "ensure that the voice of the people who depend on the NHS is heard and acted upon" (Department of Health, 1999b, p3). Ways to enable patients and the public to become informed decision-makers about health and healthcare were also outlined, as well as a commitment to involving patients and the public in making decisions about research (Department of Health, 1999b).

The demand in policy documents for the consideration of a wider range of evidence reflected wider debate amongst social scientists on the value of systematic reviews and EIPP (see chapter 1). Although some of these debates reflected ambivalence towards RCTs for evaluating HP&PH interventions, the common ground was a demand for systematic reviews to include contextual detail provided by 'qualitative' research to help inform the development, implementation and applicability of interventions. Within this climate the DH brief for the work of the EPPI-Centre was to conduct reviews that went broader than 'just effectiveness'. For example in the proposal for the programme of work during the period 2001 to 2004, one of the four aims listed was "Developing methods for systematically reviewing non-trial literature, including non-randomised and 'qualitative' studies" (Oakley *et al.*, 2001).

Earlier EPPI-Centre systematic reviews in HP&PH had adapted methods developed within health care to answer questions about the effects of interventions (e.g. Harden *et al.*, 1999a; Oakley *et al.*, 1995; Oakley *et al.*, 1996; Peersman *et al.*, 1996). These early reviews (appropriately) only included outcome evaluations that studied the impact of health promotion interventions. Although 'qualitative' data (about, for example, the acceptability of interventions) were collected when they were available, the first review in the 1998 to 2001 programme offered the first

opportunity to include 'qualitative' research in a systematic and explicit way. This review – on the topic of peer-delivered health promotion for young people - asked questions about the *effects* of peer-delivered health promotion and about its *appropriateness*. Each subsequent review offered further opportunities to include 'qualitative' research.

b) Systematic reviews

From 1998 onwards, the HP&PH work at the EPPI-Centre began to include 'qualitative' research alongside trials in systematic reviews. The seven EPPI-Centre systematic reviews in HP&PH that are relevant to this thesis are listed in table 4.1.

Table 4.1: EPPI-Centre systematic reviews in HP&PH used as sources of data in two new methodological studies on quality in 'qualitative' research

Bibliographic details of review*	**Short title**
1) Harden A, Weston R, Oakley A (1999) *A Review of The Effectiveness and Appropriateness of Peer-Delivered Health Promotion for Young People.*	**Peer-delivered health promotion**
2) Harden A, Rees R, Shepherd J, Brunton G, Oliver S, Oakley A (2001) *Young People and Mental Health: A systematic review of barriers and facilitators.*	**Young people and mental health**
3) Rees R, Harden A, Shepherd J, Brunton G, Oliver S, Oakley A (2001) *Young People and Physical Activity: A systematic review of barriers and facilitators.*	**Young people and physical activity**
4) Shepherd J, Harden A, Rees R, Brunton G, Oliver S, Oakley A (2001) *Young People and Healthy Eating: A systematic review of barriers and facilitators.*	**Young people and healthy eating**
5) Brunton G, Harden A, Rees R, Kavanagh J, Oliver S, Oakley A (2003) *Children and Physical Activity: A systematic review of barriers and facilitators.*	**Children and physical activity**
6) Thomas J, Sutcliffe K, Harden A, Oakley A, Oliver S, Rees R, Brunton G, Kavanagh J (2003) *Children and Healthy Eating: A systematic review of barriers and facilitators.*	**Children and healthy eating**
7) Rees R, Kavanagh J, Burchett H, Shepherd J, Brunton G, Harden A, Thomas J, Oliver S, Oakley A (2004) *HIV Health Promotion and Men who have Sex with Men (MSM): A systematic review of research relevant to the development and implementation of effective and appropriate interventions.*	**HIV-health promotion for MSM**

*All reviews were published by the EPPI-Centre, Social Science Research Unit, Institute of Education, University of London. They are available to download at http://eppi.ioe.ac.uk/cms/

In addition to the review reports listed above, several journal articles and book chapters have reported on the substantive findings of the reviews (Brunton *et al.*, 2005; Harden *et al.*, 2001a; Rees *et al.*, 2006; Shepherd *et al.*, 2006; Thomas *et al.*, 2005) and the methods used to include 'qualitative' research in the reviews (Harden, 2006; Harden *et al.*, 2001a; 2004; Harden and Thomas, 2005; Oakley, 2004; Oliver *et al.*, 2005; Thomas *et al.*, 2004). None of these papers addressed the specific aims of this thesis. Whilst some of the papers reported and discussed the quality of the 'qualitative' research that had been included in our reviews, this thesis goes beyond these papers to review the debates relating to quality in 'qualitative' research, survey and evaluate existing tools to assess the quality of 'qualitative' research, and analyse the relationship between study quality and review findings.

An important principle of the approach we used in all seven reviews in table 4.1 was that if different types of research questions were posed, different types of studies would be required to answer them. For example, to answer the two questions in the peer-delivered health promotion review, we included two types of studies: 'outcome evaluations', designed to evaluate the effects of interventions by measuring changes in specified outcomes; and 'process evaluations' designed to examine and/or monitor the way an intervention is delivered and received (Aggleton and Moody, 1992; Tones and Tilford, 1994). Findings from process evaluations were considered to be able to help assess the appropriateness of peer-delivered health promotion by providing data on whether young people found the approach to be acceptable and on whether the approach could be implemented in the kinds of settings in which young people lead their lives.

The next three reviews listed in table 4.1 began in 1999 as part of a review series on the barriers to, and facilitators of, health and health behaviour amongst young people. Three topics were chosen - mental health, physical activity, and healthy

eating - all areas of policy priority for the DH. There was concern over a relatively high prevalence of mental health problems and suicide amongst young people, and interest in promoting physical activity and healthy eating stemmed from rising levels of obesity, poor diet and low levels of physical activity. In line with the inequalities agenda outlined earlier, all three reviews had a particular focus on socially excluded young people. At the beginning of the review series it was hypothesised that barriers and facilitators could be identified from a) 'intervention studies' (e.g. trials) distinguishing between effective, ineffective and harmful interventions and b) 'non-intervention' studies analysing factors associated with mental health, physical activity and healthy eating. A variety of research designs and methods were anticipated to make up the category 'non-intervention' studies ranging from large-scale surveys and epidemiological analyses of large datasets, to 'qualitative' studies examining people's perspectives and experiences through in-depth interviews or focus groups.

The final reports of the peer-delivered health promotion review and the reviews addressing the barriers to, and facilitators of, mental health, physical activity and healthy eating amongst young people were well received by funders and peer referees (Harden *et al.*, 1999; Harden *et al.*, 2001; Rees *et al.*, 2001; Shepherd *et al.*, 2001). Peer referees welcomed the inclusion of 'qualitative' research and some believed it had strengthened the review findings. When negotiating further reviews, the DH posed 'barriers and facilitators' questions on children and physical activity (Brunton *et al.*, 2003), children and healthy eating (Thomas *et al.*, 2003) and HIV-health promotion for men who have sex with men (MSM) (Rees *et al.*, 2004). The HIV-health promotion review was commissioned to inform the implementation of the National Strategy for Sexual Health and HIV (Department of Health, 2001b). Particular groups of interest were: younger men; men who are sero-positive for HIV; men from black and minority ethnic groups; men with lower educational

achievement; sex workers; homosexually active men who do not identify as gay or bisexual; and injecting drug users.

All seven reviews were conducted according to the standard stages of a systematic review (e.g. Cooper and Hedges, 1994; Higgins and Green, 2006): setting a well-formulated review question; establishing the scope and boundaries of the review (inclusion and exclusion criteria); developing a review protocol; searching comprehensively for studies; describing the key features of included studies; assessing their quality; and synthesising their findings. In addition to these stages, potential users of the review were involved in decision-making processes about review questions and scope via advisory or steering groups. The reviews were also conducted according to a two-stage process: (i) a descriptive mapping stage and (ii) an in-depth review stage (figure 4.1).

User involvement and a two-stage process are particular features of an EPPI-Centre review (e.g. EPPI-Centre, 2006; Peersman *et al.*, 1999; Thomas and Harden, 2003). The descriptive mapping stage is undertaken after searching and screening have been completed. Included studies are coded according to a standardised coding strategy to build up a detailed description of existing research activity relevant to answering a particular review question. The in-depth review stage moves beyond description to assess methodological quality and synthesise findings. The production of a descriptive map can facilitate further user involvement. If a large number of studies have been identified in the map, users can help to select criteria to identify a smaller set of studies for in-depth review.

At the mapping stage numbers of included studies in the seven reviews ranged from 90 (children and physical activity) to 345 (young people and mental health) (table 4.2). Compared to outcome evaluations, numbers of 'non-intervention' studies

Figure 4.1: Stages followed in seven EPPI-Centre systematic reviews in HP&PH

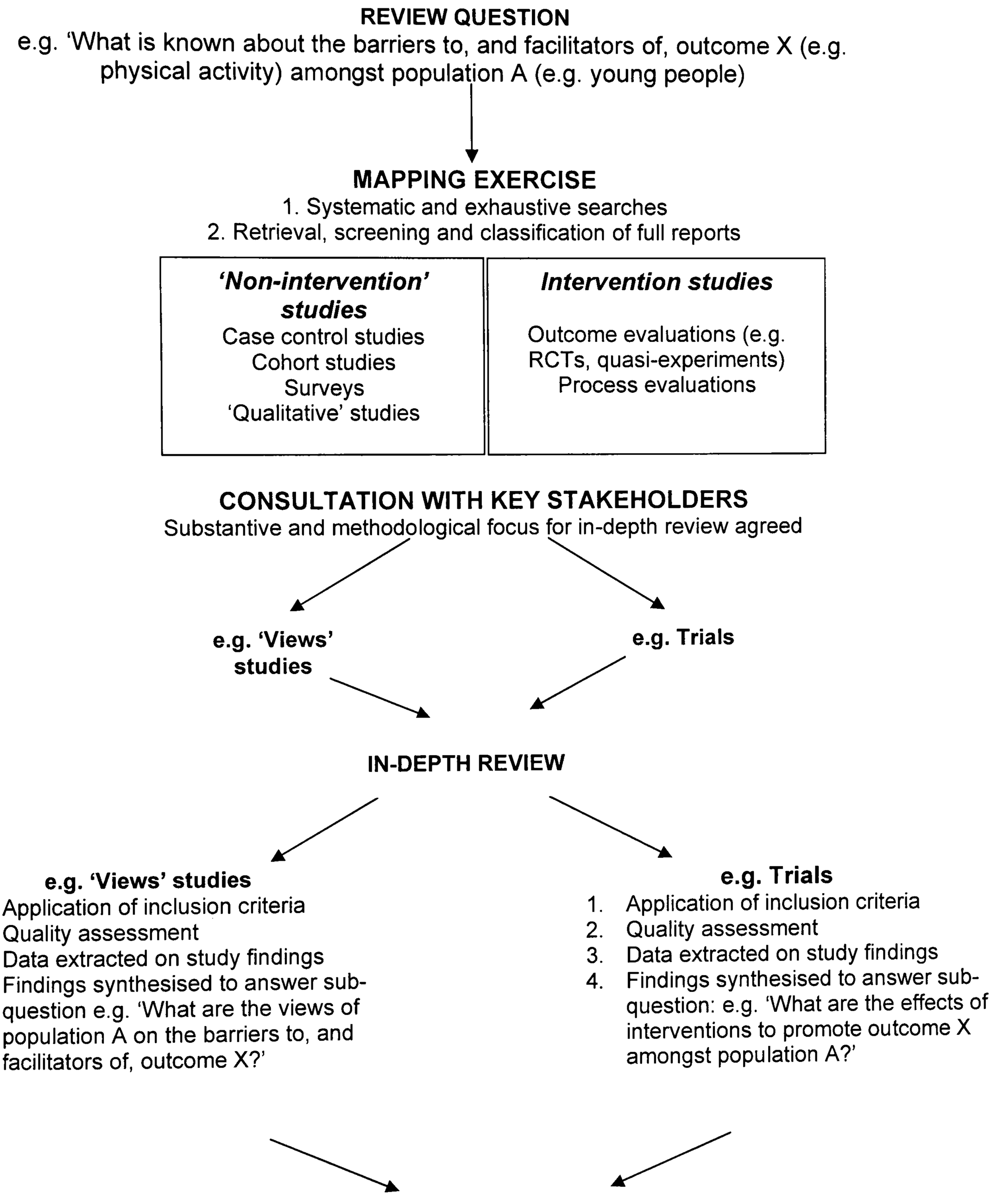


Table 4.2: Number and types of studies included in seven EPPI-Centre systematic reviews in HP&PH

	Peer-delivered health promotion	Young people and mental health	Young people and physical activity	Young people and healthy eating	Children and physical activity	Children and healthy eating	HIV-health promotion for MSM
Included in mapping exercise	**210**	**345**	**90**	**116**	**90**	**193**	**184**
Outcome only evaluations	133	149	38	64	50	113	36
Outcome evaluations with integral process evaluations	Figure not available	36	4	11	16	28	28
Process only evaluations	77	2	0	0	3	9	12
'Non-intervention' research	N/A	133	41	32	15	33	90
Systematic reviews	N/A	25	7	9	6	10	18
Included in in-depth review	**65**	**33**	**28**	**30**	**26**	**41**	**26**
Outcome only evaluations	31	6	3	9	14	12	6
Outcome evaluations with integral process evaluations	18	8	9	13	7	21	6
Process only evaluations	16	N/A	N/A	N/A	N/A	N/A	N/A
Studies of people's perspectives and experiences	N/A	12	16	8	5	8	14
Systematic reviews	N/A	7	N/A	N/A	N/A	N/A	N/A

tended to be smaller and this reflected review inclusion criteria. The review team, in consultation with its advisory group, sought international outcome evaluations with no time limit on their publication date but restricted their searches for non-intervention studies to those conducted in the UK and published in 1990 or after. Since all reviews aimed to inform current policy and practice in the UK, it was argued that the findings of non-intervention research would help to describe the specific contextual factors in the UK that influenced the health problem under review.

Within each review a smaller sub-set of studies, chosen in consultation with the review team advisory group, were subjected to more in-depth analysis. While the substantive criteria for the selection of the sub-sets varied from review to review, all of the reviews only analysed in-depth a) outcome evaluations that showed quality markers (e.g. employment of control or comparison group) and b) non-intervention studies that had examined participant's own perspectives on the health problem under review ('views' studies for short). Each review was therefore faced with two different types of studies to analyse.

The process for in-depth review for each study type followed the same basic steps of screening, data extraction, quality assessment and synthesis (figure 4.1), but the tools and methods used to conduct each one varied. For example, both sets of studies had their own clearly defined inclusion criteria; they were assessed for their quality according to standards for their specific study types; different protocols were used to extract data about their findings; and different methods were used to synthesise their findings. All reviews had three syntheses. The first synthesis pooled the effect sizes from trials using narrative synthesis, supplemented in two reviews with statistical meta-analysis. The second synthesis 'aggregated', rather than pooled the findings of studies examining people's perspectives and experiences.

Findings were broken down, interrogated and then combined into a whole via a listing of themes. The generation of 'theories' derived from people's perspectives and experiences about which interventions might work is another way of conceptualising the products of this type of synthesis (Harden *et al.*, 2004).

The third synthesis has been described as a 'cross-study' synthesis (Oliver *et al.*, 2005) or a 'mixed methods' synthesis (Harden and Thomas, 2005; Thomas *et al.*, 2004). The implications for interventions derived from people's own perspectives about the health issue under study were juxtaposed against the interventions evaluated by trials. This was done using a matrix that facilitates a comparative analysis moving back and forth between the products of the 'views' synthesis and the narrative descriptions of the interventions provided in trial reports. Three questions guided this analysis: 'which interventions match recommendations derived from people's views and experiences?'; 'which recommendations have yet to be addressed by soundly evaluated interventions?'; and 'do those interventions which match recommendations show bigger effect sizes and/or explain heterogeneity?'. Matches, mismatches and gaps were identified with gaps being used as a basis for recommending what kinds of interventions need to be developed and tested in the future.

The data underpinning all of the systematic reviews discussed in this section were stored on EPPI-Reviewer a specialist web-based research synthesis tool developed at the EPPI-Centre (Thomas, 2002). EPPI-Reviewer facilitates the collection and storage of three types of data from each study included in a systematic review: free-text data (single words or narrative); categorical data; and numeric data. For analysis, the data on EPPI-Reviewer are held in a powerful SQL database. This enables quick, sophisticated and sensitive searches to be performed. Frequencies, cross-tabulations and summary reports of categorical and free-text data aid

description and presentation of data, and a meta-analysis function is provided for the synthesis of statistical data. The original version of EPPI-Reviewer was developed in 1993 and it has undergone several major revisions as a result of the needs of the HP&PH work and other programmes of work in the EPPI-Centre. EPPI-Reviewer was very important for this thesis. It was a key factor in my decision to use EPPI-Centre reviews rather than reviews conducted externally. Use of EPPI-Reviewer meant that the studies and data underpinning EPPI-Centre reviews were stored electronically in a way that facilitated re-analysis for the methodological work I conducted in study three. I also used EPPI-Reviewer to help store and analyse the data collected for my survey and evaluation of tools to assess the quality of 'qualitative' research in study one.

4.4 EPPI-Centre programme of work in education

The English Department for Education and Skills (DfES) awarded the EPPI-Centre a five-year contract in 2000 to become a co-ordinating centre in evidence-informed policy and practice in education. The establishment of such a centre was recommended by a review of educational research in Britain as part of a strategy to overcome the problems that it had identified with the quality, relevance and accessibility of research (Hillage *et al.*, 1998). The invitation to tender document (which was issued in 1999 from the (then) Department for Education and Employment) specified the "development of arrangements comparable to the Cochrane Collaboration to prepare, maintain and promote the accessibility of systematic reviews of research relating to policies and practices in education" via the development of "international research review groups". However, the tender document also specified the need to look beyond RCTs in recognition that "much high quality educational research employs a qualitative or combined quantitative and qualitative approach". Accordingly, the centre proposed by the EPPI-Centre

aimed to: develop a number of review groups (RGs); train and support RGs to conduct systematic reviews; develop and make available methods and tools for the systematic review of different study designs, including qualitative studies and RCTs; and develop a web-based system for undertaking, storing, disseminating and updating reviews. It was proposed that all of this would be undertaken within a "framework committed to transparency, responsiveness to criticism and equity of access"[1].

This five-year programme of work at the EPPI-Centre represented the first major UK resource for systematic review work exclusively within the field of education. The DfES continues to fund this work and more recently the EPPI-Centre has been funded by the English Teacher Development Agency (TDA) to support additional education reviews and RGs. In contrast to the programme of work on HP&PH at the EPPI-Centre, the DfES funding covers the costs of developing systematic review capacity and the centre infrastructure rather than specific review products themselves. For example, in the original five year programme of work funded by the DfES, registered RGs were given 'seed' monies to support them in their preparation of systematic reviews but the scope of individual RGs and their reviews were not pre-specified by the DfES.

To support RGs, we developed a systematic review handbook (EPPI-Centre, 2006); produced guidelines for data extraction and quality assessment of educational research (EPPI-Centre, 2002; Gough, 2004; Gough, forthcoming); and further developed EPPI-Reviewer (Thomas, 2002). To date around 30 groups have registered with the EPPI-Centre to conduct reviews of research in education.

---

[1] This quote is reproduced from the EPPI-Centre proposal (led by Ann Oakley and David Gough) to the English Department for Education and Employment to become a centre for evidence-informed policy and practice in education.

Examples of past and present RGs are: Art and Design; Assessment and Learning; Citizenship; Continuing Professional Development; Early Years; English Teaching; Modern Foreign Languages; and Personal, Social and Health Education.

The work of the EPPI-Centre RGs has thrown up many of the same political, technical and conceptual challenges encountered in conducting reviews in HP&PH and other areas (Gough and Elbourne, 2002; Oakley, 2003; 2006; Oakley *et al.*, 2005). For example, there were difficulties in finding research on electronic databases; low yields of useable studies; and difficulties around judging quality, especially in small fields in which researchers all knew each other. Another problem for RGs was what to do with the large numbers of 'qualitative' studies. The first four EPPI-Centre systematic reviews in education all addressed impact questions in one form or another, asking about the effect of some intervention on outcomes (e.g. Dyson *et al.*, Francis *et al.*, 2002). RGs had problems dealing with the 'qualitative' studies using procedures based on systematic reviews for answering questions about the average balance of benefit and harm from interventions. The findings of this thesis are therefore relevant to meeting a major challenge faced by those attempting to review educational research in a systematic way. They are also relevant in other areas where researchers take this approach (e.g. Fisher, 2005; Wallace *et al.*, 2004).

4.6 Summary and conclusion

In this chapter I have given a brief description of the aims, design and methods of the three new methodological studies in my thesis and a description of relevant EPPI-Centre programmes of work in which my thesis originates. The descriptions show how these programmes generated the data that I used in my thesis: the analysis of the development of the new quality assessment tool for 'qualitative' research in study two was a product of both the HP&PH and education work, and

the systematic reviews that were analysed in study three were a product of the HP&PH work. The description of EPPI-Centre programmes of work also locates the methodological questions in my thesis within larger efforts to advance the science of systematic reviews to improve the quality and relevance of the evidence-base to inform decision-making in public policy. My thesis explores one particular aspect - the problem of including 'qualitative' research in systematic reviews - within this larger agenda to incorporate a more diverse range of research into the evidence-base.

This chapter has also shown how my thesis complements rather than duplicates existing or past EPPI-Centre programmes of work. I described how both the education and the HP&PH work had a remit to incorporate diverse types of evidence into systematic reviews, and I also described how the systematic reviews produced in the HP&PH programme included 'qualitative' research. However, neither of these programmes had the specific aims addressed in this thesis concerning the inclusion of 'qualitative' research in systematic reviews. For example, neither programme of work aimed to review the conceptual debates in the literature, or to survey and evaluate existing tools developed to assess the quality of 'qualitative' research, or to analyse the relationship between study quality and review findings. The work that I have undertaken for my thesis therefore represents an original contribution within the EPPI-Centre as well as within the wider research literature.

CHAPTER 5

Study 1: A survey and evaluation of existing tools for assessing the quality of 'qualitative' research

5.1 Introduction

This chapter reports the first of three new methodological studies to advance knowledge about how to include and assess the quality of 'qualitative' research in systematic reviews. Despite debates about assessing quality amongst 'qualitative' researchers (see chapter three), a lot of energy has been devoted to developing sets of quality criteria or questions against which reports of 'qualitative' studies can be assessed in the form of 'users' guides'; 'critical appraisal tools' or 'guidelines for peer referees'. With the rise of evidence-informed policy and practice and systematic reviews, interest in using these tools to identify good quality evidence from 'qualitative' research has grown. Systematic reviewers, for example, are interested in these tools because of their potential to be used within the quality assessment stage of a systematic review (e.g. Kahn *et al.*, 2001).

Several recent overviews have compared some of these tools or the concept of 'checklists' more generally. Some of this work has highlighted the potential dangers of using tools in a rigid or prescriptive way especially given the lack of consensus on quality in 'qualitative' research (Barbour, 2001; Chapple and Rodgers, 1998). Other overviews have focused on describing the kinds of tools that exist (Angen, 2000; Devers, 1999; Eakin and Mykhalovskiy, 2003; Katrak *et al.*, 2004; Madill *et al.*, 2000; Murphy *et al.*, 1999; Spencer *et al.*, 2003). A major argument from this work is that tools differ according to the philosophical position of tool authors on the nature

of knowledge and knowledge production in 'qualitative' research. Tools have been classified into one of two main categories, with the second generally seen as the more valuable. The first category is variously described as 'positivist', 'naïve realist', or 'empirically based', and the second as 'post-positivist', 'interpretivist', or 'philosophically based'. For example, Spencer *et al.* (2003, p95), who characterised tools as either 'philosophically based' or 'empirically based', argued that 'empirically based' tools (i.e. those tools which do not attend explicitly to the "ontological or epistemological base of 'qualitative' research" in their development) rarely mentioned "paradigm specific" features such as 'subjective meaning', 'reflexivity', 'saturation', 'context', 'thick description', or 'negative case analysis'.

Despite this body of work, there has been no systematic attempt to identify, describe, and evaluate existing tools with a view to drawing out the specific implications for their use in systematic reviews. This is in contrast to the work that has been done to provide systematic overviews of tools for assessing the quality of trials (e.g. Juni *et al.*, 1999; Moher *et al.*, 1995). This work has documented the variation and commonalities across tools and their strengths and weaknesses (e.g. does a particular tool contain items to assess the presence of design flaws which have been empirically or theoretically related to the occurrence of bias in effect sizes?; has the tool been tested?). The results of this work therefore provide a useful resource for reviewers. One of the aims of the study reported in this chapter is to provide an equivalent resource that systematic reviewers can draw upon to inform decisions about how they might assess the quality of 'qualitative' studies in their review. Finding good tools to assess the quality of 'qualitative' studies is also important as an end in itself. Guidance on which tools are useful could benefit a number of groups including researchers for guiding the conduct and write-up of 'qualitative' research; funders for commissioning research and evaluating end-of

award reports; and peer reviewers. Useful tools may help to drive up standards in the conduct and reporting of 'qualitative' research.

The study described in this chapter aimed to search systematically for reports relevant to the topic of assessing the quality of 'qualitative' research, and to identify critical appraisal tools to apply to study reports. Three questions were posed i) what kinds of tools exist for assessing the quality of 'qualitative' studies?; ii) to what extent do these tools differ and to what extent do they overlap?; and iii) what are the strengths and limitations of the tools and which might be useful tools for systematic reviews?

5.2 Methods

The study was carried out in four stages: (i) identification of tools; (ii) collection of data from tools; (iii) analysis of tools; and (iv) evaluation of tools.

(i) Identification of tools

Several sources of published and unpublished literature were searched, grouped into two overall strands as described below. All citations identified were downloaded or entered onto a reference management database. Titles and abstracts were scanned for relevance and full reports obtained. Full reports were assessed for inclusion to identify those that described a structured approach to quality assessment (e.g. a set of questions or prompts) for application to reports of 'qualitative' studies ('tools').

a) 'General' bibliographic databases

Six bibliographic databases known to index literature relevant to the social sciences were searched (ASSIA, ERIC, MEDLINE, the Social Science Citation Index, Sociological Abstracts and Social Services Abstracts). Searches were conducted from the inception of each database to March 2002. Free-text terms for "qualitative" research (e.g. 'qualitative', 'ethnography'; 'interpretative'; 'focus groups') were combined with free-text terms for 'quality' (e.g. 'standards', 'validity', 'trustworthiness'). These combinations of terms were sought in the titles of the bibliographic citations held on the databases.

b) Other sources

Searches on the above sources were supplemented by four other strategies, implemented throughout 2002. (I stopped searching for tools at the end of 2002). Firstly, four specialist databases known to index methodological research related to systematic reviews were searched (the Cochrane Methodology Database, held on the Cochrane Library; the ESRC Evidence Network database, held by Queen Mary and Westfield College; the UK Health Technology Assessment Database; and the database prepared by the proposed Cochrane Collaboration's Qualitative Methods Group and the Campbell Collaboration's Process and Implementation Methods Group). For the Cochrane Methodology database, a simple search strategy was employed to identify the term 'qualitative' in any field of the bibliographic citations held on the registers. For the other three registers, *all* bibliographic citations held were scanned for relevance. Secondly, a simple search of the web was undertaken using the phrases 'quality of qualitative research' and 'trustworthiness of qualitative research' on the GOOGLE search engine. Thirdly, the methodological papers identified in the course of programmes of systematic review work in health promotion and education at the EPPI-Centre were sifted. Finally, as full reports were screened for inclusion, the reference lists of those meeting the inclusion criteria were scanned to identify further potentially relevant citations.

These ways of searching were particularly useful for identifying reports which did not have a focus on the quality assessment of 'qualitative' research reflected in their title or abstract, but were nevertheless relevant to the review (e.g. systematic reviews which had included 'qualitative' research and assessed its quality).

(ii) Collection of data from tools

Data were collected from each tool using a standardised form designed specifically for this study. This data collection form, which is presented in full in Appendix A, covered the following four sections:

a) Identifying details

This section covered the country(ies) in which tool author(s) was/were based; the discipline or professional background of the author(s); the year in which the tool was published/reported; the format in which the tool was published/reported; and whether there was any funding to support the development of the tool.

b) Conceptual underpinnings and tool development

This section covered the definition of 'qualitative' research given/used by the tool author(s); where tool author(s) locate themselves within the debates about assessing the quality of 'qualitative' research; the reasons the author(s) gave for why they developed the tool; and how the tool was developed.

c) Tool content and structure

This section covered whether the tool was developed for use with a particular type of 'qualitative' research or within a particular discipline or applied field of study; the number and content of items in the tool; and the nature and type of any guidance

provided by the author(s) on how to use the tool (e.g. how to use the tool to make an overall judgement on the quality of a particular study).

d) Evaluation

This section covered whether the tool had been evaluated (e.g. by applying it to several reports of 'qualitative' research). If the tool had been evaluated details about who evaluated it and the results of the evaluation were recorded.

(iii) Analysis of tools

All data were held and analysed using a specialist web-based reviewing programme (EPPI-Reviewer) (Thomas, 2002). Questions were mainly open ended and the data generated were mainly in free-text form. For example, in response to the question 'What reasons does/do the author(s) give for why they developed the tool?' the relevant parts of the tool report were copied verbatim into the reviewing programme. When pre-defined categories were used to code answers, the data were analysed using simple counts. The answers to open-ended questions were analysed using methods for qualitative data.

Procedures associated with content analysis (Kippendorf, 2004, Mayring, 2000) were used to characterise, for example, the range of reasons authors offered for why they developed their tool. The text within the relevant answer was listed against each tool. Each line of text was examined and a list of reasons generated. Reasons within this list were grouped, if appropriate, to form wider abstract categories. For example the reasons 'for a systematic review'; 'to assess the quality of 'qualitative' research in medical journals' were grouped into the wider category 'to use in a review of qualitative research'. This final list of categories was used to code the free-text data.

When free-text answers were more extensive (e.g. for tool content), the data were exported to NVivo for more detailed sorting and analysis. Open coding techniques, as suggested by texts outlining methods for analysis of qualitative data (e.g. Miles and Huberman, 1994), were used to describe the type and content of items contained within and across tools. The diversity of tool content was difficult to capture. A number of different coding schemes were tried out before settling on the one actually used. For example, one scheme that was tried and rejected allocated items into one of three categories: how well the study was reported; how well the study was carried out; and the quality of the findings. The problem with this scheme was that it was often difficult to distinguish between items assessing the quality of reporting versus items assessing how well the study was carried out. In addition there was no provision in this scheme for assessment of research questions or of the literature or theory which framed the study.

The final coding scheme classified tool items into one of five domains: background literature, theory and research questions; sampling, sample and setting; fieldwork, data collection and analysis; findings; and write-up and ethics. To compare tools according to the weight they gave to each of the five domains, the number of tool items falling into each domain was calculated as a proportion of the total number of items in that tool. A 'spider-graph'[1] was prepared for each tool to display the distribution of items across the five domains. Spider-graphs are a type of visual graphic that can be used to aid the analysis of multi-dimensional data (Chambers *et al.*, 1983). They are constructed by assigning each data dimension to a separate axis. In this study, spider graphs had five axes labelled A to E (A: background literature, theory and research questions; B: sampling, sample and setting; C: fieldwork, data collection and analysis; D: findings; and E: write-up and ethics). The

[1] Spider-graphs are also called 'star-plots', 'footprint graphs' or 'radar plots'.

proportion of items within different domains was plotted on these five axes for each tool.

(iv) Evaluation of tools

Each tool was assessed against ten desirable features of a quality assessment tool to apply to 'qualitative' research in systematic reviews. In order to get a wider perspective beyond my own views on the qualities of a useful tool, these ten features were derived from the views and experiences of six researchers with experience of conducting systematic reviews from one university department[2]. Although I could have involved researchers from other universities to maximise diversity in perspective, I chose to involve researchers from my own university department to take advantage of the convenience this offered in terms of recruitment and meeting organisation. I felt that the most important issue was that I did not rely solely on my own perspective. I did, however, attempt to introduce diversity in perspective by asking a researcher who I did not work closely with day to day (RS) and a researcher who, at that time, had only just joined the university (MN).

To elicit their views each researcher was asked to apply one or two of the tools identified to one or two reports of 'qualitative' studies. Eight different tools, chosen to represent different types of tools, were applied by the researchers: Cesario *et al.* (2002); Giacomini and Cook (2000a,b); Long and Godfrey (2004); Mays and Pope (2000); Popay *et al.* (1998); Sandelowski and Barroso (2002); Spencer *et al.* (2003);

---

[2]The researchers were Mark Newman (MN), Ann Oakley (AO), Sandy Oliver (SO), Rebecca Rees (RR), Ruth Stewart (RS) and James Thomas (JT). All researchers were from the Social Science Research Unit at the Institute of Education, University of London. AO and SO supervised this PhD. Although AO and SO went beyond their supervisory roles for this part of study two, I remained solely responsible for the design, conduct and write-up of the whole study.

Whittemore *et al.* (2001). Five different reports of 'qualitative' research were used, again chosen to represent different types of studies. Types of studies represented were: a study using in-depth interviews to investigate the effects of social support on recidivism amongst male prisoners (Breese *et al.*, 2000); a grounded theory study of reflective practice in research supervision for doctoral theses (Douglas, 2003); a study evaluating an assisted self-help group for drug users via interviews and observation (Felix-Ortiz *et al.*, 2000); a survey of the views of people with learning difficulties with respect to work and employment support via semi-structured interviews (Wistow and Schneider, 2003); and a study exploring explanations of teenage pregnancy and motherhood given in in-depth interviews with teenage mothers and health professionals (Arai, 2003).

The researchers were invited to attend a meeting to discuss their experiences of using the tools. For those unable to attend the meeting feedback was given via e-mail or phone. A list of problems and strengths within the tools were recorded and these were transformed into 10 desirable features of useful tools.

5.3 Results

(i) Identification of tools, bibliographic details and tool development

a) Identification of tools

A considerable amount of energy has gone into developing tools for assessing the quality of 'qualitative' studies; the searching and screening process resulted in the identification of 31 tools (table 5.1). Initially, searches of multiple sources resulted in 244 citations judged to be relevant on the basis of their title and abstract (if available). Full reports were obtained for 216 of these. On inspection of the full

Table 5.1: Tools to assess the quality of 'qualitative' research (N=31): bibliographic details, tool authors, tool purpose, key quality concepts, number of items, and focus of tool items

Tool citation	Tool authors	Purpose	Key quality concepts	Items	Focus of tool items
Beck (1993)	Academic nurse	To help readers appreciate "the scientific merit of [qualitative] research studies" (p254)	Credibility, fittingness and auditability	26	Methods orientated
Blaxter (1996)	Sociologists	To provide guidelines for 'medically orientated' journal editors	Rigour and sophistication	20	Methods orientated
Boulton and Fitzpatrick (1997)	Sociologists	To provide "basic advice" (p83) to readers and users (unfamiliar with qualitative evidence) within the health service to evaluate qualitative papers.	No explicitly stated beyond 'explicit standards and principles for good practice'	10	Methods orientated
Boulton *et al.* (1996)	Sociologists	To use in a methodological review of qualitative studies to raise questions about the "direction" of this type of research within healthcare (p172)	No explicitly stated beyond 'explicit guidelines and standards'	18	Methods orientated
Britten *et al.* (1995)	Sociologists Academic doctors	To help those working in general practice who want to know more about qualitative research	Not explicitly stated	11	Findings orientated
Campbell *et al.* (2003)	Sociologists	To use in a meta-ethnography of qualitative studies	Not explicitly stated	15	Methods and findings orientated
CASP (1998)	Sociologists Academic doctors	To help those working in health and social care to appraise qualitative evidence and put knowledge into practice.	Not explicitly stated	10	Methods and findings orientated
CASP (2002)	Qualitative researchers Research users	To help those working in health and social care to appraise qualitative evidence and put knowledge into practice.	Not explicitly stated	10	Methods and findings orientated
Cesario *et al.* (2002)	Academic nurses	To use in the critical appraisal of qualitative research relevant to the development of clinical guidelines for nursing management in the second stage of labour	Descriptive vividness; methodological congruence; analytical preciseness; theoretical connectedness; heuristic relevance	43	Methods and findings orientated
Corbin and Strauss (1990)	Sociologists	To provide evaluative guidance to readers of grounded theory publications and to provide systematic guidelines for authors of grounded theories	Adequacy of the research process and empirical grounding of findings	14	Methods and findings orientated
Drisko (1997)	Academic social worker	To help social work journal editors and peer referees to fairly evaluate manuscripts reporting qualitative studies	Academic integrity	6	Methods and findings orientated

Table 5.1 (continued): Tools to assess the quality of 'qualitative' research (N=31): bibliographic details, tool authors, tool purpose, key quality concepts, number of items, and focus of tool items

Tool citation	Author discipline	Purpose	Key quality concepts	Items	Focus of tool items
Elder and Miller (1995)	Sociologists Academic doctors	To help family physicians, who might be unfamiliar with qualitative research, to read and assess qualitative research to inform their clinical practice.	Trustworthiness, believability and contribution	16	Methods orientated
Elliot *et al.* (1999)	Psychologists	To provide psychology journal editors and peer referees with "guidance on how to conduct appropriate reviews of qualitative research manuscripts" (p215)	No explicitly stated beyond 'standards for good practice'	14	Findings orientated
Forchuck and Roberts (1993)	Academic nurses	To help those who might be unfamiliar with qualitative research to read it with a critical eye (e.g. student nurses, health professional unfamiliar with qualitative methods)	Not explicitly stated	7	Methods and findings orientated
Giacomini and Cook (2000a,b)	Sociologists Academic doctors	To help doctors use qualitative research to improve the care they provided to patients.	Systematic observation; competent interpretation, correspondence with social reality experienced by participants; meaningful to users of the research	8	Methods and findings orientated
Greenhalgh and Taylor (1997)	Academic doctors	To help those working in healthcare who want to use qualitative research to inform their practice	No explicitly stated beyond mention of tool provides 'ground rules'	9	Methods and findings orientated
Hoddinott and Pill (1997)	Sociologists Academic doctors	To use in a methodological review to assess studies using in-depth interviews	Detailed methodological information for replication to confirm findings	8	Methods orientated
Kuzel and Engel (2001)	Academic nurses	To help health professionals to "evaluate qualitative inquiry with attention how well it informs practice" (p117)	Plausible explanations and findings that can inform practice	6	Methods and findings orientated
Long and Godfrey (2004)	Sociologists	To use in a systematic review in the area of health and social care	Systematic methods, rigour, appropriateness, production of accurate accounts	34	Methods orientated
Malterud (2001)	Academic doctors	To provide guidelines for authors and journal editors for preparing or assessing a manuscript reporting a qualitative study	Validity, relevance and reflexivity	30	Methods orientated
Mays and Pope (1995)	Sociologists	To give readers of research in healthcare confidence in the critical appraisal of qualitative research	Explicit reporting; plausible and coherent account; generalisability	11	Methods orientated

Table 5.1 (continued): Tools to assess the quality of 'qualitative' research (N=31): bibliographic details, tool authors, tool purpose, key quality concepts, number of items, and focus of tool items

Author	Author discipline	Purpose	Key quality concepts	Items	
Mays and Pope (2000)	Sociologists	To provide users and funders with a means to assess the quality of qualitative research.	Validity and relevance	14	Methods and findings orientated
McLaughlin (1986)	Sociologists	To help anyone involved in the evaluation of qualitative research (e.g. journal editors, other researchers)	Reliability, validity, reducing analytical bias, credibility, empirical grounding, plausibility	8	Findings orientated
Miles and Huberman (1994)	Sociologists	For researchers to apply to their own qualitative research or for readers of qualitative research	Objectivity/ confirmability; reliability/dependability/auditability; internal validity/ credibility/ authenticity; external validity/ transferability/ fittingness; and utilisation/ application/ action orientation.	50	Findings orientated
Muecke (1994)	Academic nurses	To show the richness, value and strengths of ethnography	Not explicitly stated	6	Findings orientated
Popay *et al.* (1998)	Sociologists	To help those who want to use qualitative research in systematic reviews, especially those who might be unfamiliar with this type of research.	Interpretation of subjective meaning, description of social context, and attention to lay knowledge	7	Findings orientated
Sandelowski & Barroso (2002)	Academic nurses	To use in a meta-synthesis of qualitative studies	Aesthetic and rhetorical criteria as well as epistemic criteria	81	Methods and findings orientated
Spencer *et al.* (2003)	Sociologists	To a) drive up standards of qualitative evaluation and b) to help government assess the worth of research for informing policy.	Contribution, defensible in design, rigorous in conduct, credible in claim.	18	Findings orientated
Treloar *et al.* (2000)	Sociologists Academic doctors	To help researchers to develop the skills to critically appraise and carry out qualitative research	Empirical grounding, rigour, minimise or make bias explicit, produce theoretical and generalisable understanding	10	Methods orientated
Vermeire *et al.* (2003)	Academic doctors	To help medical journal editors, peer referees and authors critically appraise and prepare manuscripts which report studies using focus groups	Validity, relevance and transparency	12	Findings orientated
Whittemore *et al.* (2001)	Academic nurses	To facilitate "the decision-making process for investigators and the evaluative process for consumers of research." (p535).	Credibility, authenticity, criticality, and integrity	10	Findings orientated

report, 47 turned out not to be about assessing the quality of 'qualitative' research and were excluded. A further 129 reports were about this topic but they did not describe a tool. These reports included theoretical discussions, review articles and editorials. Nine reports were published in languages other than English but no attempt was made to get these reports translated. Although the extra effort and resources required for translation would not have been great, there was no evidence to suggest that reports not published in English a) actually described tools or b) provided a different perspective or approach to quality not already covered by the 31 tools published in English. English abstracts were provided for five of the non-English reports and these abstracts suggested discussions about the quality of 'qualitative' research rather than tools.

It was not easy to locate the 31 tools. Only eight of the tools were identified via the major bibliographic databases searched. Despite searching a range of databases that index social science research, the most productive database was MEDLINE on which six of the eight tools were found. A further 18 tools were found via scanning the reference lists of relevant reports. Of the remaining five tools, one was identified on the Cochrane Methodology Database and four were identified via opportunistic personal contact with tool authors.

b) Bibliographic details

The earliest date a tool was published was 1986 (McLaughlin, 1986) and the latest was 2004 (Long and Godfrey, 2004)[3]. The majority of tools were developed by authors based in the UK (n=15) or the US (n=12). Two were developed by authors based in Canada (Forchuck and Roberts, 1993; Giacomini and Cook, 2000a,b) and one each in Belgium (Vermeire *et al.*, 2002) and Denmark (Malterud, 2001). The UK

[3] Even though I had found all tools by the end of 2002, some of these were not published until after this date (Campbell *et al.*, 2003; Long and Godfrey, 2004; Spencer *et al.*, 2003).

and North American dominance may reflect an English Language bias in the databases and other sources searched.

All but six tools were published in journal articles. Three were published as chapters within books on 'qualitative' research (Kuzel and Engel, 2001; Miles and Huberman, 1994; Muecke, 1994), and three were published as stand-alone reports (CASP, 1998; CASP, 2002; Spencer *et al.*, 2003). The 25 reports of tools in journal articles were published in a total of 21 journals. Five of the tools were published in three of the world's leading medical journals (British Medical Journal; Journal of the American Medical Association and the Lancet). Nursing and general practice journals had also published a number of tools (e.g. Western Journal of Nursing Research, the Journal of Family Practice). Other tools were published in methodological journals (e.g. International Journal of Qualitative Methods, Qualitative Health Research) or social science journals (e.g. Social Science and Medicine, British Journal of Clinical Psychology, Qualitative Sociology).

c) Tool development and testing

The majority of tools were developed to apply to healthcare research (n=22) and this is likely to reflect the more longstanding interest in using research to inform policy and practice in health compared to others areas. Two tools had been developed to apply to research relevant to social work or social care (Drisko, 1997; Long and Godfrey, 2004), and one tool had been developed in each of the following areas: education (McLaughlin, 1986), psychology (Elliot *et al.*, 1999) and medical sociology (Blaxter, 1996). Four tools did not specify a particular area or discipline (Corbin and Strauss, 1990; Miles and Huberman, 1994; Spencer *et al.*, 2003; Whittemore, 2001).

The tools were developed by authors within a variety of disciplines (table 5.1). The majority of tools were developed by sociologists (Blaxter, 1996; Boulton and Fitzpatrick, 1997; Boulton *et al.*, 1996; Campbell *et al.*, 2003; Corbin and Strauss, 1990; Long and Godfrey, 2004; Mays and Pope, 1995; 1998; McLaughlin, 1986; Miles and Huberman, 1994; Popay *et al.*, 1998; Spencer *et al.*, 2003); academic nurses (Beck, 1993, Cesario *et al.*, 2002; Forchuck and Roberts, 1993; Kuzel and Engel, 2001; Muecke, 1994; Sandelowski and Barroso, 2002; Whittemore *et al.*, 2001); academic doctors (Greenhalgh and Taylor, 1997; Malterud, 2001; Vermeire *et al.*, 2003); and collaborations between sociologists, academic doctors and/or epidemiologists (Britten *et al.*, 1995; CASP, 1998; Elder and Miller, 1995; Giacomini and Cook, 2000a,b; Hodinott and Pill, 1997; Treloar *et al.*, 2000). One tool was developed by psychologists (Elliot *et al.*, 1999) and one tool was developed by academic social workers (Drisko, 1997). The remaining tool was described as the result of collaboration between a group of 'qualitative' researchers and research users (CASP, 2002).

There was limited information on how tools were developed but many tools were based on previous tools or theoretical discussions about quality in 'qualitative' research. A small number of tools were developed with the input of a wider group of people that just the tool authors (Blaxter, 1996; Elliot *et al.*, 1999; Sandelowski and Barroso, 2002; Spencer *et al.*, 2003). Preliminary versions of the tools described by Blaxter (1996) and Elliot *et al.* (1999) were revised in the light of feedback from (respectively) participants at a Medical Sociology Conference and a meeting of the Society for Psychotherapy Research; Sandelowski and Barroso (2002) asked a group of 'qualitative' synthesis experts to try out a preliminary version of their tool; and Spencer *et al.* (2003) sought feedback from representatives of several groups on a preliminary version of their tool (policy-makers, research commissioners and funders, managers, and academics who conduct 'qualitative' research or write about

quality in 'qualitative' research). The authors for only four tools explicitly reported that they were developed with the aid of specific funding (Long and Godfrey, 2004; Popay *et al.*, 1998; Sandelowski and Barroso, 2002; Spencer *et al.*, 2003).

Only nine of the 31 tools had been tried out or evaluated. For three of these nine tools, the authors simply stated that the tool had been tried out and did not report any detail on how the tool worked (Cesario *et al.*, 2002; Popay *et al.*, 1998; Spencer *et al.*, 2003). For the other six tools, evaluations covered levels of agreement between researchers' appraisal of study reports using the tool and/or reflections on the most useful and least useful items in the tools (Boulton *et al.*, 1996; Campbell *et al.*, 2003; Hoddinott and Pill, 1997; Mays and Pope, 1995[4]; Sandelowski and Barroso, 2002; Vermeire *et al.*, 2002).

(ii) Tool purpose, type of 'qualitative' research and key quality concepts

a) Tool purpose

Tool authors offered a variety of purposes for their tools (table 5.1). There were at least seven different reasons why tools were developed and sometimes tool authors offered more than one reason:

1. To help journal editors, peer referees and authors to assess or prepare a manuscript reporting a 'qualitative' study for publication (Blaxter, 1996; Drisko, 1997; Elliot *et al.*, 1999; Malterud, 2001; McLaughlin, 1986; Vermeire *et al.*, 2003).

[4] An evaluation of this tool is reported in O'Conner *et al.* (2001).

2. To use in reviews of 'qualitative' research, either methodological reviews (Boulton *et al.*, 1996; Hoddinott and Pill, 1997) or systematic reviews in a substantive area (Campbell *et al.*, 2003; Long and Godfrey, 2004; Sandelowski and Barroso, 2002).

3. To critically appraise 'qualitative' research before using it to inform clinical guidelines (Cesario *et al.*, 2002).

4. To help those who use, fund or read 'qualitative' research to evaluate this type of research (Boulton and Fitzpatrick, 1997; Britten *et al.* 1995; Forchuck and Roberts, 1993; Mays and Pope, 1995; Mays and Pope, 1998; Treloar, 2000)

5. To help practitioners and policy-makers appraise 'qualitative' research with a view to using that research to inform their practice (CASP, 1998; CASP, 2002: Elder and Miller, 1995; Giacomini and Cook, 2000a,b; Greenhalgh and Taylor, 1997; Kuzel and Engel, 2001; Spencer *et al.*, 2003; Whittemore *et al.*, 2001).
6. To provide guidelines for researchers conducting 'qualitative' research (Corbin and Strauss, 1990; Miles and Huberman, 1994; Whittemore *et al.*, 2001); and

7. To help readers appreciate the scientific nature of 'qualitative' studies (Beck, 1993) or to reveal the strengths of 'qualitative' research (Cesario *et al.*, 2002; Muecke, 1986).

In describing their tools, authors raised other issues which cut across the seven purposes outline above. Several tool authors emphasised that their tool would be especially suited to helping those unfamiliar with 'qualitative' research (Blaxter, 1996 Boulton and Fitzpatrick, 1997; Britten *et al.*, 1995; CASP, 1998; CASP, 2002; Drisko, 1997; Elder and Miller, 1995; Forchuck and Roberts, 1993; Kuzel and Engel,

2001; Popay *et al.*, 1998; Treloar *et al.*, 2000; Vermeire *et al.* 2003) because they are, for example, a student or a researcher who has been "raised in another research tradition" (Kuzel and Engel, 2001, p115). Other tool authors saw a role for their tool in driving up the quality of 'qualitative' research in the future (Elliot *et al.*, 1999; Long and Godfrey, 2004; Spencer *et al.*, 2003). For example, Elliot *et al.* (1999, p218) believe that a rapid increase in 'qualitative' research has resulted in poorly executed or "no method" research and guidelines for evaluating 'qualitative' research will encourage better quality control.

Some tool authors highlighted the importance of their tool for ensuring that 'qualitative' research would be judged by appropriate standards. Without specific tools, authors feared that "traditional scientific criteria relevant to quantitative studies" might be used (Beck, 1993, p265) and as a consequence 'qualitative' research may be "misunderstood and judged inferior" (Popay *et al.*, 1998). Tool authors argued that appropriate tools need to be "grounded in the qualitative paradigm" (Long and Godfrey, 2004, p194) or be consistent with the epistemology of 'qualitative' research (Drisko, 1997). On a related theme was a concern to legitimise 'qualitative' research and its contribution to knowledge. Authors argued that providing a tool to help readers assess quality would, for example, demonstrate the existence of methodological guidance and standards of rigour for 'qualitative' research (Blaxter, 1996; Elliot *et al.*, 1999); prevent assessments of 'qualitative' studies as "hopelessly subjective and unscientific" (Elder and Miller, 1995, p279); and enable readers to "capture the richness and depth of qualitative findings" (Cesario *et al.*, 2002, p713).

<u>b) Types of 'qualitative' research</u>

Although some tool authors recognised that what goes on under the name of 'qualitative' is diverse, all but seven designed their tool for application to 'qualitative

research' not further specified. The authors of two tools stated that their tools were designed for evaluations (Long and Godfrey, 2004; Spencer *et al.*, 2003). A further two tools were designed for ethnography (McLaughlin, 1987; Muecke, 1994), one for focus groups (Vermeire *et al.* 2002), one for grounded theory (Corbin and Strauss, 1990) and one for studies using in-depth interviews (Hodinott and Pill, 1997).

For those tool authors who offered definitions, 'qualitative' research was defined according to its purpose, distinctive features, and typical methods of data collection and analysis. Examples of the different purposes described were: to understand, interpret or illuminate the subjective meanings shaping action and behaviour; to understand the dynamics of social life; to understand the relationship between process and outcome; to build theory; and to provide evidence on the appropriateness of interventions. Examples of distinctive features were: the collection of 'rich' data; the use of unstructured, flexible and/or sensitive methods; a focus on natural setting; a 'holistic' perspective; the role of the researcher as part of the research; and an inductive approach.

Some tool authors highlighted the kinds of questions they thought that 'qualitative' research could not answer but that 'quantitative' research could (e.g. 'how much' questions, testing hypothesised relationships, predicting outcomes). This suggests a view of 'qualitative' and 'quantitative' research as complementary, and the majority of tool authors either explicitly or implicitly adopted such a view. A small number of tool authors appeared to suggest that a 'qualitative' approach was 'better' than a 'quantitative' one. Greenhalgh and Taylor (1997, p740) suggested that 'qualitative' researchers seek "a deeper truth"; Sandelowski and Barroso (2002, p2) suggest that 'qualitative' researchers have a "general distaste for and distrust of 'mainstream' research"; and Cesario *et al.* (2002, p713) argue that 'qualitative'

researchers approach "situations from a worldview that is comprehensive and holistic, rather than reductionistic and deterministic".

c) Key quality concepts

A range of quality concepts were used by tool authors with most saying something about what standards or dimensions of quality their tool would help to assess 'qualitative' research against (table 5.1). A number of tool authors located themselves within the debate about whether 'qualitative' research should be judged against the same criteria as 'quantitative' research or whether alternative criteria should be used. For example Beck (1993, p264) viewed reliability and validity as inappropriate for 'qualitative' research and designed her tool to assess "credibility, fittingness and auditability" whereas Mays and Pope (1998, p50) argued that reliability and validity could be applied if they were tailored to reflect the "distinctive goals of qualitative research".

Despite this debate, many tool authors used both 'traditional' and 'alternative' quality concepts within the same tool. For example, Miles and Huberman (1994) organised the items in their tool under five dimensions of quality i) objectivity/ confirmability; ii) reliability/dependability/auditability; iii) internal validity/credibility/authenticity; iv) external validity/ transferability/ fittingness; and v) utilisation/ application/ action orientation. Some or all of these five dimensions of quality were apparent in all tools with the fifth dimension about whether findings are useful and relevant being a particular feature of those tools designed to help research users appraise 'qualitative' research to inform their practice.

Regardless of the terms used to describe dimensions of quality, tool authors offered some very similar ideas about what constitutes 'validity' and 'credibility'. Mays and Pope (1995, p110) defined 'validity' as the production of a plausible and coherent

explanation of the phenomenon under study and Giacomini and Cook (2000a, p479) define it as whether "the analysis offers a meaningful approximation to the truth of social phenomenon". Using the same language as Mays and Pope (1995), Blaxter (1996) suggested that credibility should be assessed according to whether a plausible and coherent account had been produced. For Whittemore *et al* (2001, p 527) assessing 'credibility' means asking whether interpretations are accurate and whether they "reveal some truth external to the investigators experience". Long and Godfrey (2004, p180) argued that we need to assess studies according to whether they produce "plausible accounts that reflect what is being examined in as accurate a way as possible".

Implicit here are ideas about reducing distortion, bias and error. These same ideas were embodied in terms such as 'systematic', 'rigorous', and 'explicit', which were often used within tools. Only a third (n=11) of tools used the terms 'bias'. There was no agreement, however, amongst those tool authors who used the term 'bias' about what should be done with it. Several authors argued that we should strive to reduce, eliminate, limit, prevent, and/or protect against bias (Beck, 1993; Cesario *et al.*, 2002; Corbin and Strauss, 1990; Mays and Pope, 1995; Treloar *et al.*, 2000; Whittemore *et al.* 2001). For example Treloar *et al.* (2000, p350) suggest that key to the "qualitative application" of concepts like validity and credibility is "the move to minimise the effects of bias on data collection, analysis and interpretation". Other tool authors argued that we can only declare our biases (Drisko, 1997; Elder and Miller, 1995; Elliot *et al.*, 1999; Greenhalgh and Taylor, 1997; Popay *et al.*, 1998). For example, Elder and Miller (1995, p280) suggest that the researcher "becomes part of the research and describes rather than eliminates known biases" and Greenhalgh and Taylor (1997, p 742) argue that "the most" that can be expected of researchers with respect to bias is that they "describe in detail where they are coming from so that the results can be interpreted accordingly."

Another dimension of quality represented in the tools was the 'appropriateness' or the fit between the epistemological approach, the methods used and the research objectives. For example, a number of tools opened with a question about whether a 'qualitative' approach was appropriate to address the research question (e.g. Blaxter, 1996; Britten *et al.*, 1995; Malterud, 2001). In other tools, the fit between the epistemological approach, the methods used and the research objectives was a major theme running throughout (e.g. Drisko, 1997; Popay *et al.*, 1998; Sandelowski and Barroso, 2002). In the tool designed by Popay *et al.* (1998) a 'primary marker' of quality is whether or not the study had adopted a 'verstehen' approach to knowledge in which primacy is given to the way people within particular groups, cultures and societies view the phenomenon under study. 'Secondary markers' assess whether particular aspects of the study (e.g. sampling, data collection) are consistent with the primary marker.

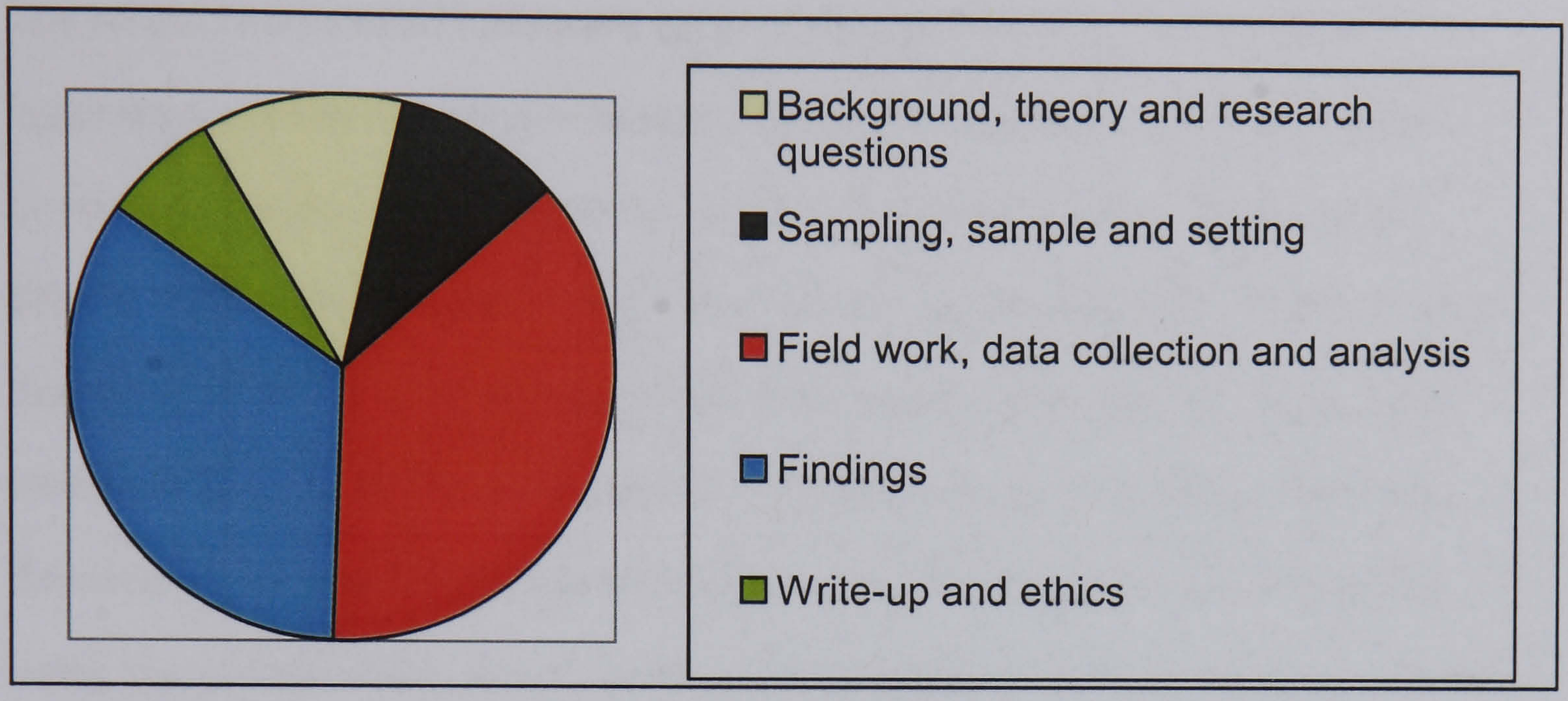
(iii) Overview of tool content

Tools varied in length, with the number of items on which a judgement was required ranging from 6 to 81 (table 5.1). Across the tools, a total of 515 different items to assess the quality of 'qualitative' research were offered. There were only 24 duplicate items and the most common duplicate appearing in 10 of the 31 tools was 'Is a qualitative methodology appropriate to address the aims of the study?'[5]

[5] Other duplicate items were: the data analysis was not sufficiently rigorous; the relationship between researchers and participants has not been adequately considered; there is no clear statement of findings; the findings of this study are not transferable to a wider population; there is no clear statement of the aims of the research; the sampling strategy is not appropriate to address the aims; the data were not collected in a way that addressed the research issue; informed consent was not obtained from the participants and documented; and the analysis of the data was not systematic.

In the analysis of tool content I allocated each different item into one of five areas depending on which domain of a study the item invited judgement on and/or where readers might have to look in a study report to make their judgement (figure 5.1).

Figure 5.1: Distribution of tool items according to five study domains (N=519)



The two largest groups of items involve making judgements based on an examination of the fieldwork, data collection and analysis (n=191, 37%) or the findings of the study (n =178, 34%). There were three smaller groups of items covering the background to the study, the literature review, any theoretical framework used and research questions (n=62, 12%); sampling, sample and setting (n=53, 10%); and write-up and ethics (n=38, 7%). More details about the content of the tools within each of these domains are described in the next section.

Each tool differed according to the proportion of items it contained in each of the five domains discussed above (figure 5.2). Although each tool was unique, there were some common patterns in the shapes of their 'spider-graphs' (the name of the figure used to display the proportion of tool items in each of the five areas). The tools in figure 5.2 are listed according to the shape of their spider-graphs, with similar shapes grouped together. There were three distinct groups: i) tools which have a preponderance of items about fieldwork, data collection and analysis ('methods-

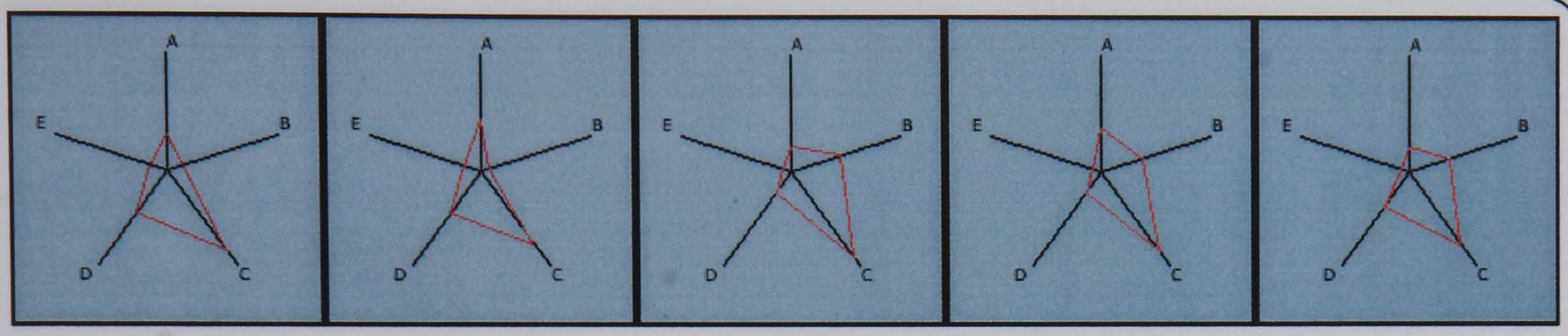
orientated tools') (n=10); ii) tools with a predominance of items which invite judgements about the findings of the study ('findings-orientated tools') (n=9); iii) tools with the majority of items spread evenly across both methods and findings ('methods- and findings-orientated tools') (n=12).

The *methods-orientated tools* were either a) designed for those unfamiliar with 'qualitative' research working in healthcare such as healthcare managers, family physicians, medical journal editors or readers of medical journals (Beck, 1993; Blaxter, 1996; Elder and Miller, 1995; Boulton and Fitzpatrick, 1997; Malterud, 2001; Treloar *et al.*, 2000); or b) for use in methodological reviews (Boulton *et al.*, 1996; Hoddinott and Pill, 1997). An exception to this was Long and Godfrey (2004) who developed their tool for use in a systematic review of research in social care. By using one of these tools, reviewers would be prompted to make assessments about whether or not to rely on a study's findings for a review largely on the basis of the methods used in the study: whether they were systematic and rigorous; explicitly reported; and appropriate given the research question.

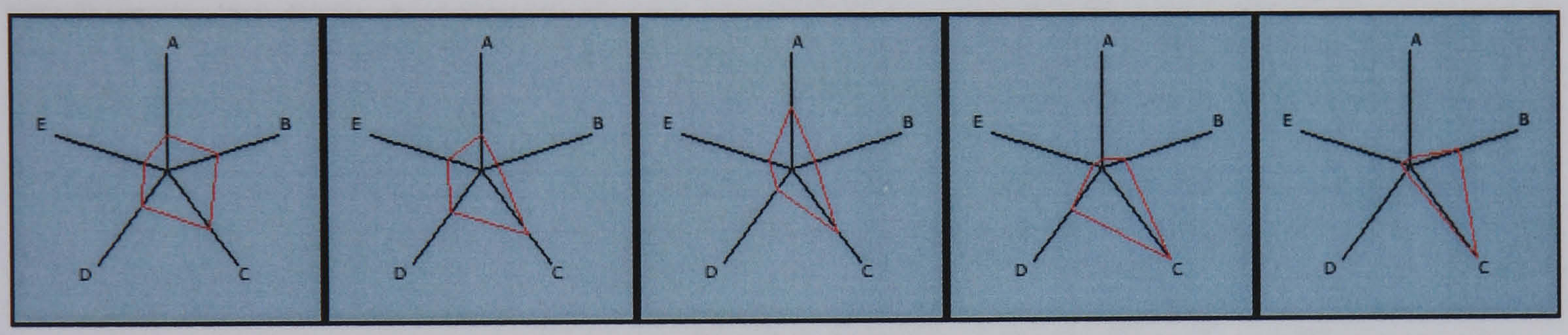
In contrast to the intended 'unfamiliar' audiences of the methods orientated tools, *findings-orientated tools* appear to have been developed mainly for more specialist audiences. With the exception of Vermeire *et al.* (2003) and Britten *et al.* (1995), these tools were published in journals or books with a focus on 'qualitative' methods (Miles and Huberman, 1994; Muecke, 1994; Popay *et al.*, 1998; Whittemore *et al.*, 2001) or in social science journals (Elliot *et al.*, 1999; McLaughlin, 1986). The tool designed by Spencer *et al.* (2003) was designed for government researchers to help in the conduct and appraisal of 'qualitative' evaluations on policy questions. By using one of these tools reviewers would be required to give more attention to the findings of studies when making a judgement about quality.

Figure 5.2: Configuration of tool items across five domains* in tools to assess the quality of 'qualitative' studies (N=31)

*Key to domains
A: Background, theory and research questions;
B: Sampling, sample and setting
C: Methods of data collection and analysis
D: Findings
E: Write-up and ethics

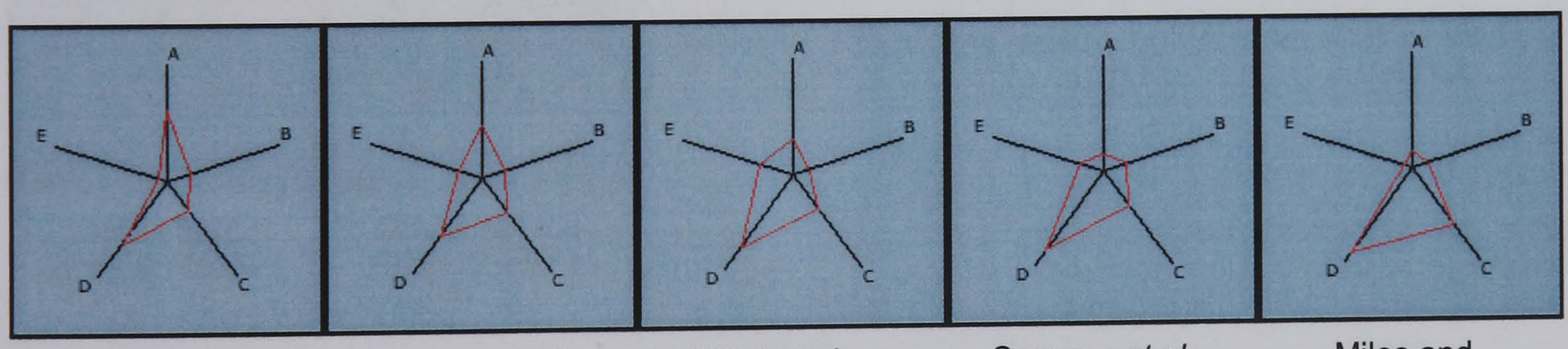


Blaxter (1996) | Elder and Miller (1995) | Mays and Pope (1995) | Boulton and Fitzpatrick (1997) | Boulton *et al.* (1996)

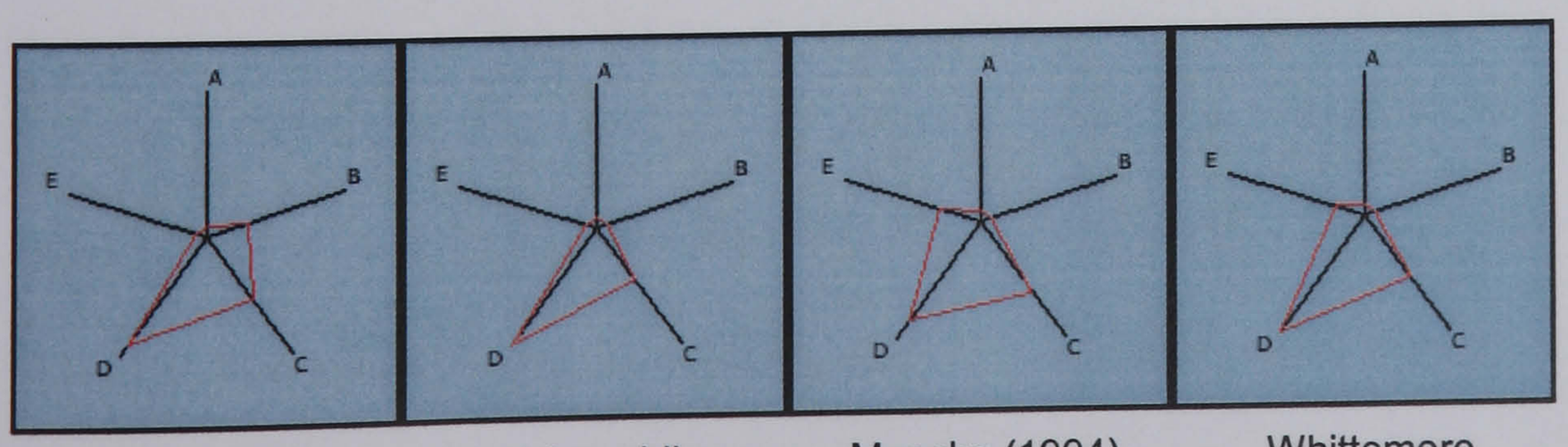


Long and Godfrey (2004) | Malterud (2001) | Treloar *et al.* (2000) | Beck (1993) | Hoddinott and Pill (1997)

Methods-orientated



Vermeire *et al.* (2003) | Britten *et al.* (1995) | Elliot *et al.* (1999) | Spencer *et al.* (2003) | Miles and Huberman (1994)

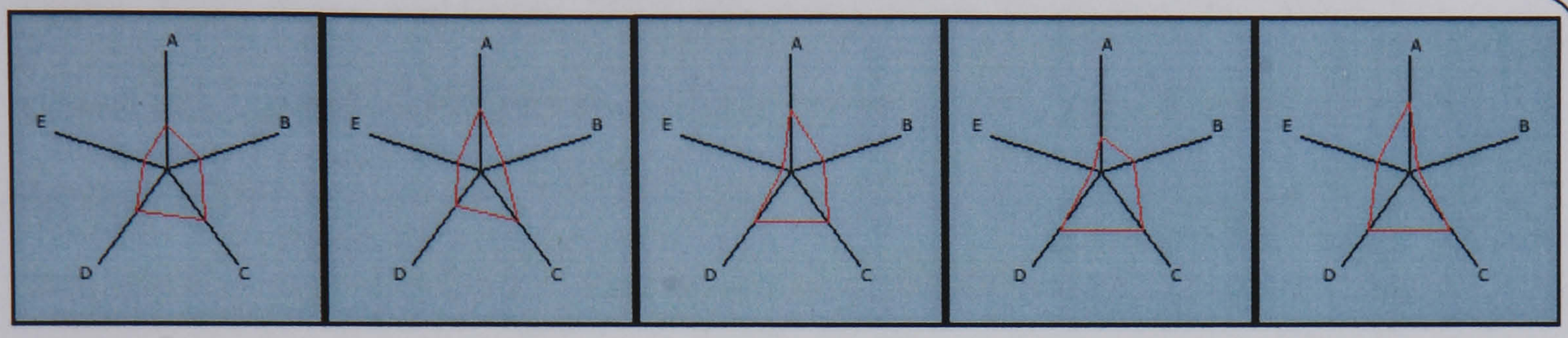


Popay *et al.* (1998) | McLaughlin (1986) | Muecke (1994) | Whittemore *et al.* (2001)

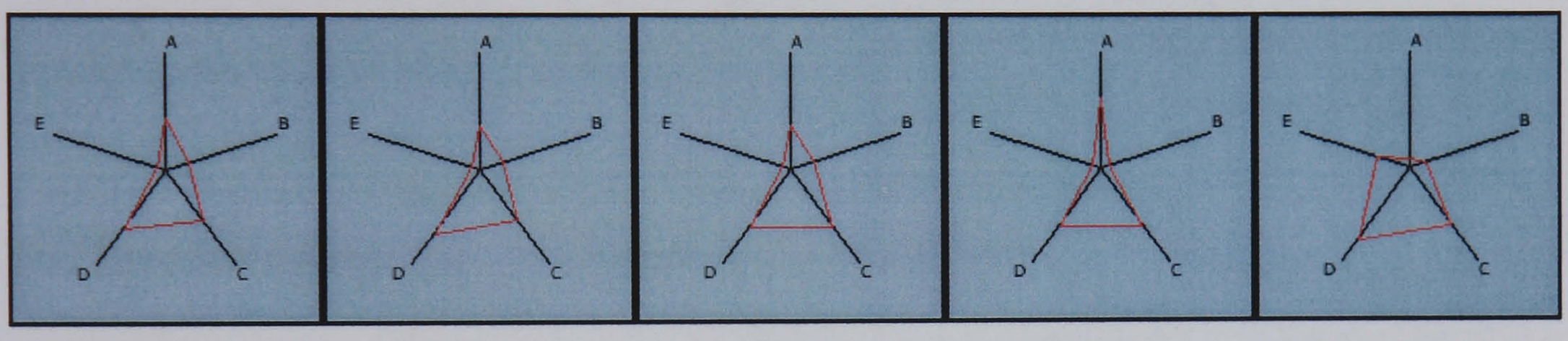
Findings-orientated

Figure 5.2 (continued): Configuration of tool items across five domains* in tools to assess the quality of 'qualitative' studies (N=31)

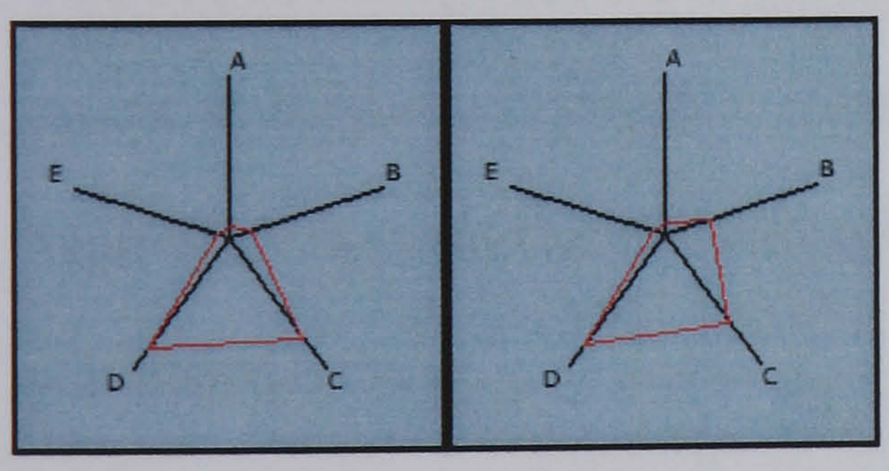
*Key to domains
A: Background, theory and research questions;
B: Sampling, sample and setting
C: Methods of data collection and analysis
D: Findings
E: Write-up and ethics



Sandelowski and Barroso (2002) | CASP (2002) | Forchuck and Roberts (1993) | Mays and Pope (2000) | Drisko (1997)



Campbell *et al.* (2003) | CASP (1998) | Greenhalgh and Taylor (1997) | Kuzel and Engel (2001) | Cesario *et al.* (2002)



Corbin and Strauss (1990) | Giacomini and Cook (2000a,b)

Methods- and findings-orientated

The *methods- and findings-orientated tools* had similar purposes to the methods-orientated tools. The majority were designed either for a) those unfamiliar with 'qualitative' research working in health or social care such as healthcare managers, doctors, nurses, or social workers (CASP, 1998; CASP, 2002; Drisko, 1997; Forchuck and Roberts, 1993; Giacomini and Cook, 2000a,b; Greenhalgh and Taylor, 1997; Kuzel and Engel, 2001; Mays and Pope, 1995) or b) for use in systematic reviews (Campbell *et al.*, 2003; Cesario *et al.*, 2002; Sandelowski and Barroso 2002). The exception to this was the tool by Corbin and Strauss (1990). This tool had more in common with the findings-orientated tools because it was designed for authors or readers of grounded theories and published in *Qualitative Sociology*. However, this tool gave equal attention to the adequacy of the research process and the empirical grounding of findings.

(iv) Detailed description and analysis of tool content

This section describes the items in the tools in more detail within each of the five domains shown in Figure 5.1. Differences in tool content were sought between 'methods-orientated tools', 'findings-orientated tools' and 'methods- and findings-orientated tools'.

<u>a) Background literature, theory, research questions, and design</u>

Individual tool items in this domain invited judgements on whether (number of tools containing items on this issue):

- The study is located within existing knowledge and/or this location is appropriate and adequate (n=6).
- A theoretical framework/perspective is specified and/or is reflected in the way the study was carried out (n=8).

- The question or phenomenon under study is an important and/or relevant one (n=5).
- The aims/research questions are stated and/or clearly formulated (n=16).
- A 'qualitative' approach was appropriate to address the research questions/ an appropriate rationale was provided for using 'qualitative' methods (n=18).

Tools that covered this domain tended to be those that required an appraisal of every aspect of studies and their report such as those designed for journal editors or those which took unfamiliar readers step by step through a study report. Those that did not cover this domain were those focused on particular dimensions of quality or stages of research (Beck, 1993; Cesario *et al.*, 2002; Corbin and Strauss, 1990; Giacomini and Cook, 2000a,b; Hoddinott and Pill, 1997; Popay *et al.* 1998; McLaughlin, 1986; Muecke, 1994; Whittemore *et al.*, 2001). For example Corbin and Strauss (1990) aimed to assess the adequacy of the research process and the empirical grounding of findings within studies using grounded theory and Hoddinott and Pill (1997) aimed to assess the methods in studies using in-depth interviews. Regardless of purpose, methods-orientated and methods- and findings-orientated tools tended to dominate in this domain, particularly for the first three issues listed above. With respect to the first of these three issues, tools asked whether the study was connected to existing knowledge and whether this connection was clear (Blaxter, 1996; Long and Godfrey, 2004) or whether there was a clear scientific context for the study (Elliot *et al.*, 1999; Forchuck and Roberts, 1993). Two tools asked for more detailed information about the literature review such as whether key studies are included and/or whether the review shows a clear logic and critical attitude (Malterud, 2001; Sandelowski and Barroso, 2002). Regarding theory, tools asked whether a theoretical framework had been used and clearly identified (Blaxter, 1996; Campbell *et al.*, 2003; Drisko, 1997; Mays and Pope, 1995; Miles

and Huberman, 1994; Long and Godfrey, 2004; Sandelowski and Barroso, 2002; Treloar *et al.*, 2000); whether the theoretical framework was in fact reflected in the way the study was carried out (Long and Godfrey, 2004; Sandelowski and Barroso, 2002); whether the theoretical framework fitted the phenomenon under study (Sandelowski and Barroso, 2002); whether the framework influenced the researchers before or after they went into the field (Sandelowski and Barroso, 2002); or whether study authors were sensitive to links between theories, values and facts (Kuzel and Engel, 2001).

Regarding the question under study, tools asked whether the study was worth doing (Britten *et al.*, 1995; Greenhalgh and Taylor, 1997; Malterud, 2001; Mays and Pope, 2000); or whether the authors demonstrated that the study was worth doing (Sandelowski and Barroso, 2002). Half of the tools asked whether the aims of the study were clearly stated (Boulton and Fitzpatrick, 1997; Boulton *et al.*, 1996; Britten *et al.*, 1995; Campbell *et al.*, 2003; CASP, 1998; CASP, 2002; Drisko, 1997; Elder and Miller, 1995; Forchuck and Roberts, 1993; Greenhalgh and Taylor, 1997; Long and Godfrey, 2004; Malterud, 2001; Miles and Huberman, 1994; Sandelowski and Barroso, 2002; Treloar *et al.*, 2000; Vermeire *et al.*, 2003). Nearly all of these tools, and three others, went on to ask whether the methods used were an appropriate choice given the research question. The majority of these tools asked whether 'qualitative' design/methods were appropriate (Blaxter, 1996; Boulton and Fitzpatrick, 1997; Boulton *et al.*, 1996; Britten *et al.*, 1995; Campbell *et al.*, 2003; CASP, 1998; CASP, 2002; Greenhalgh and Taylor, 1997; Malterud, 2001; Mays and Pope, 2000; Sandelowski and Barroso, 2002; Treloar *et al.*, 2000; Vermeire *et al.*, 2003). Others simply asked whether the design/methods were appropriate given the research question (Drisko, 1997; Elder and Miller, 1995; Forchuck and Roberts, 1993; Miles and Huberman, 1994; Spencer *et al.*, 2003).

b) Sampling, sample and setting

Individual tool items in this domain invited judgements on whether (number of tools containing items on this issue):

- The sampling strategy was described and/or described adequately (n=8).
- The sampling strategy was appropriate given, for example, the research question or aims of the study (n=14).
- The characteristics of the sample and/or setting were described and/or described adequately (n=11).
- The sample and/or setting was adequate or appropriate (n=6).

Five tools did not cover issues of sampling and sample and all of these where either findings-orientated or methods- and findings-orientated tools. However, three of these five tools did cover sampling issues indirectly via items about the generalisability of findings (Drisko, 1997; Kuzel and Engel, 2001; McLaughlin, 1986). It was not clear why the other two tools ignored sampling issues, although both of these were findings orientated tools (Muecke, 1994; Whittemore *et al.*, 2001). When findings-orientated tools did include items about sampling and sample these focused on only two of the four issues listed above: whether the sampling strategy was appropriate and whether the sample was described adequately. With respect to the appropriateness of the sampling strategy, some tools simply asked whether the strategy was appropriate given the research question (Campbell *et al.*, 2003; CASP, 1998; CASP, 2002; Cesario *et al.*, 2002; Treloar *et al.*, 2000; Vermeire *et al.*, 2003) or whether the sampling strategy was justified or well reasoned (Blaxter, 1996; Britten *et al.*, 1995; Giacomini and Cook, 2000a,b; Mays and Pope, 1995; Spencer *et al.*, 2003). Other tools asked whether theoretical or purposeful sampling had been used (Popay *et al.*, 1998; Sandelowski and Barroso,

2002). Two tools asked whether the sampling was theoretically diverse enough to encourage broader applicability (Mays and Pope, 1995; Miles and Huberman, 1994).

With respect to whether the sample was described adequately, some tools simply asked whether sample characteristics had been described (Beck, 1993; Boulton and Fitzpatrick, 1997; Forchuck and Roberts, 1993). Others asked whether detailed profiles had been given (Spencer *et al.*, 2003); whether key characteristics had been presented (Long and Godfrey, 2004); whether enough information had been provided for the reader to be able to relate the findings to other groups or settings (Elliot *et al.*, 1999; Mays and Pope, 2000; Miles and Huberman, 1994); or whether detail on characteristics critical to the understanding of the study context or findings had been given (Malterud, 2001; Sandelowski and Barroso, 2002). One tool specified that age, gender, ethnicity, social class and other relevant demographic characteristics should be described (Boulton *et al.*, 1996).

Methods-orientated and methods- and findings-orientated tools included items about whether the sampling strategy had been described adequately and whether the final sample was adequate or appropriate. With respect to the former issue, tools wanted to know how study samples had been selected and recruited (Beck, 1993; Boulton and Fitzpatrick, 1997; Boulton *et al.*, 1996; Corbin and Strauss, 1990; Greenhalgh and Taylor, 1997; Hodinott and Pill, 1997; Long and Godfrey, 2004; Mays and Pope, 1995). With respect to the latter issue, various markers of adequate or appropriate were specified. Tools asked whether sample size and configuration had sufficient depth and width (Long and Godfrey, 2004); included a range of informants experiencing the phenomenon under study to support conceptual rather than statistical generalisations (Beck, 1993; Boulton *et al.*, 1996; Mays and Pope, 2000); were relevant to the research questions (Giancommini and

Cook, 2000a,b); or could support theoretical saturation, the holistic study of particulars and the findings (Sandelowski and Barroso, 2002). Sandelowski and Barroso (2002) also asked whether the sample size could support the study findings.

c) Fieldwork, data collection and analysis

All tools contained items in this domain. Collectively, tool items invited judgements on whether (number of tools containing items on this issue):

- Methods of data collection and analysis were described/described adequately (n=18).
- Methods of data collection and analysis were appropriate given, for example, the research question/aims (n=14).
- Strategies were used and/or appropriate strategies were used in data collection and/or data analysis to minimise bias, distortion or error (n=23).
- Distortion, bias or error were introduced into the study (n=5).
- There is consideration/adequate consideration of the role of the researcher, their relationship to participants, and possible effects on the research (n=16).

Examples of each type of tool were represented in each of the five areas listed above. For example, tools of each type demanded explicit detail on methods of data collection and analysis (e.g. Blaxter, 1996; Sandelowski and Barroso, 2003; Spencer *et al.*, 2003) and tools of each type raised similar issues on whether data collection strategies were appropriate such as evidence of a flexible and sensitive approach (e.g. Boulton *et al.*, 1996; Mays and Pope, 2000; Popay *et al.*, 1998) or the collection of comprehensive data in terms of breadth and depth (e.g. Long and Godfrey, 2004; Spencer *et al.*, 2003; Giacomini and Cook, 2000a,b).

Tool items about whether adequate information was provided about methods of data collection and analysis ranged from the very general (e.g. was sufficient information provided on methods of data collection/data management/data analysis?) to the highly specific, which prescribed in detail what kinds of information should be provided about the methods. Methods-orientated tools mainly provided these detailed prescriptions. Collectively the tools asked for information on: where data were collected from, by whom and in what context (Boulton *et al.*, 1996; Hoddinott and Pill, 1997); how data were elicited and the range of questions asked (Boulton *et al.*, 1996; Long and Godfrey, 2004); length and timing of data collection (Long and Godfrey, 2004; Sandelowski and Barroso, 2003); the qualifications of the interviewer (Hoddinott and Pill, 1997); reasons for choice of data collection strategy (Malterud, 2001); how themes, concepts and categories were derived from the data (Blaxter, 1996; Boulton *et al.*, 1996; Corbin and Strauss, 1990; Mays and Pope, 1995); on the basis of which categories theoretical sampling proceeded (Corbin and Strauss, 1990); rules for formulation and confirmation of propositions and hypotheses (Corbin and Strauss, 1990; Miles and Huberman, 1994); the perspectives and ideas used for data analysis (Malterud, 2001); and the role of the theoretical framework in the analysis (Malterud, 2001)

Tool items about the appropriateness of methods also ranged from the general to the specific. Several tools contained items that simply asked reviewers to assess whether methods were appropriate given the purpose, research question and design of the study (Campbell *et al.*, 2003; CASP, 1998; CASP, 2002; Treloar *et al.*, 2000). Other tools contained more specific items about data collection and/or data analysis. A number of tools asked reviewers to assess whether data collection was comprehensive, flexible, and/or sensitive enough to provide a vivid, rich and/or holistic description of the phenomenon under study (Boulton and Fitzpatrick, 1997; Boulton *et al.*, 1996; Cesario *et al.*, 2002; Elder and Miller, 1995; Giacomini and

Cook, 2000a,b; Long and Godfrey, 2004; Malterud, 2001; Mays and Pope, 2000; Popay *et al.*, 1998; Sandelowski and Barroso, 2002; Spencer *et al.*, 2003). For example, Cesario *et al.* (2002) asks whether an adequate length of time was spent at the site and whether sufficient time was spent gathering data, and Elder and Miller (1995) asks whether the researchers 'keep following up' and whether there is sufficient contact between researchers and participants.

Only four tools contained specific questions about the appropriateness of data analysis. Popay *et al.* (1998) ask whether participant's perceptions and experiences are treated as knowledge in their own right and Malterud (2001) asks to what extent the analysis was guided by preconceptions rather than the data. Long and Godfrey (2004) ask whether the analysis sought breadth (contrast of two of more perspectives) as well as depth (insight into a single perspective). Sandelowski and Barroso (2002) ask a) whether the analysis is case-orientated as opposed to variable-orientated, and b) whether the analysis is done at the right level (e.g. it takes into account group interaction for the analysis of focus groups).

All but seven tools mentioned specific strategies for increasing rigour (table 5.2). Not all tools advocated the same strategies but the most popular were: searching for negative cases; checking the findings with participants; and use of multiple sources of data (triangulation). On average, methods-orientated tools advocated the use of a greater number of strategies than findings-orientated or methods- and findings-orientated tools[6]. Compared to findings-orientated tools, methods-orientated tools were more likely to advocate the use of more than one researcher to analyse data; use of consistent data collection protocols; and checking for

[6] Median number of strategies advocated by tools (range): Methods-orientated = 4 (0-6); Findings-orientated = 2 (0-5); Methods- and findings-orientated = 2 (0-4).

consistency in data analysis over time or between investigators[7]. Other strategies mentioned by the tools included the use of quantitative evidence to test 'qualitative' conclusions (Mays and Pope, 1995) and eliminating the potential for elite bias (Cesario *et al.*, 2002).

Table 5.2: Number of tools that refer to specific and commonly cited strategies for increasing rigour in 'qualitative' data collection and analysis (N=31).

Strategy	**No. of tools**
Searching for negative cases/alternative explanations	18
Checking the findings with participants	14
Use of multiple methods of data collection or data sources	12
Using more than one researcher to analyse data	10
Keeping careful records of data	7
Use of an external panel or peer review	6
Using consistent data collection protocols	3
Use of an audit trail	3
Checking for consistency in data analysis over time or between researchers	2
Use of well trained investigators	1

Seven tools did not mention any of the strategies listed in table 5.2. Apart from Hoddinott and Pill (1997), whose tool focused on how well methods are reported in in-depth interview studies, the tools which did not mention any of these strategies were either findings-orientated tools (Popay *et al.* 1998; Whittemore *et al.*, 2001) or methods- and findings-orientated tools (Corbin and Strauss, 1990; Forchuck and Roberts, 1993; Kuzel and Engel, 2001; Sandelowski and Barroso, 2002). These tools tended to either a) be formulated at a more generic or abstract level so that

[7] Elliot *et al.* (1999) was the only findings orientated tool to advocate the use of more than one researcher to analyse data. Miles and Huberman (1994) was the only findings orientated tool to advocate the use of consistent data collection protocols or checking for consistency in data analysis over time or between investigators.

individual items cut across specific aspects of the research process (e.g. Kuzel and Engel, 2001, ask whether the researchers conducting a study demonstrate a sensitivity to the linkage between presumptions, facts, values, interpretations, and theories); or b) ask open questions about rigour in data collection and analysis rather than questions about whether specific strategies were used (e.g. Popay *et al.*, 1998 ask whether there is evidence of data quality).

Half of the tools (n = 16) contained items that invited judgements about whether any consideration had been given to the impact of the researcher on the findings of the research. Items emphasised either consideration of the effect of the researcher's background, values, or consideration of the effect of the relationship between the researcher and the participants. These items, which bring into play the concept of reflexivity, were a feature of methods-orientated tools as well as findings-orientated tools. Interestingly, these items featured in tools whose authors believed that 'qualitative' researchers should strive to reduce the effects of bias (e.g. Beck, 1993; Mays and Pope, 2000) as well as in those tools whose authors believed that researchers can only hope to declare biases. Furthermore, strategies associated with the reduction of bias, distortion and error, such as those detailed in table 5.2, were advocated by tool authors who argued that researchers can only hope to declare biases rather than reduce them (e.g. Drisko, 1997; Elder and Miller, 1995).

<u>d) Findings</u>

All but one of the tools included items about the findings of the study (Hoddinott and Pill, 1997). Individual tool items in this domain invited judgements on whether (number of tools containing items on this issue):

- Findings are clear, coherent and distinguishable (n=5).
- Findings addressed the aims of the study (n=5).

- Concepts, ideas and/or theory are well developed (n=12).
- Descriptions are detailed and context-rich (n=9).
- Findings are grounded in/supported by the data (n=22).
- Findings are stable across data sources, contexts, time, and or/researchers (n=2).
- Findings illuminate and/or illuminate in a believable way subjective meaning and participant perspectives and experiences, and diversity in meaning, perspectives, and experiences is explored well (n=3).
- Readers view the findings as meaningful and can recognise and/or understand the experiences/phenomenon described (n=8).
- Participants view the findings as accurate and/or an honest and caring description of their experiences (n=3).
- Wider inferences can be drawn from the study findings (to other contexts, settings and groups of people)/study findings can be assessed according to whether wider inferences can be made (n=14).
- Findings are compared to and/or are congruent with previous empirical or theoretical work (n=10).
- Findings do not make a contribution to knowledge and/or are not useful for practice (n=18).

The items about findings within methods-orientated tools asked about a more limited set of the above issues than either findings-orientated or methods- and findings-orientated tools. In the main, methods-orientated tools only asked about those aspects of study findings which were assessed by the largest number of tools: whether findings were grounded in the data; whether wider inferences can be drawn

from the study findings; and/or whether findings make a useful contribution to knowledge and/or policy and practice[8].

Whether study findings were supported by data (i.e. field notes, extracts from interviews) was the aspect of study findings assessed by the largest number of tools (N=22). Highlighting study findings as interpretations, these tools wanted to see sufficient data presented to convince the reader that interpretations are valid, credible or trustworthy (Boulton and Fitzpatrick, 1997; Britten *et al.*, 1995; Greenhalgh and Taylor, 1997; Spencer *et al.*, 2003; Vermeire *et al.* 2003); to support the relationship between evidence and conclusions (CASP, 2002; Long and Godfrey, 2004; Mays and Pope, 1995; Treloar *et al.*, 2000); to show correspondence or a fit between the data and interpretations of the data (Beck, 1993; Cesario *et al.*, 2002; Elliot *et al.*, 1999; Forchuck and Roberts, 1993; Mays and Pope, 1995; Mays and Pope, 2000); to support claims of recurrent patterns or theory development (Beck, 1993; Drisko, 1997; Malterud, 2001; Miles and Huberman, 1994); to help readers judge the range of evidence being used/see how the researcher's arrived at their findings (Beck, 1993; Blaxter, 1996; Miles and Huberman, 1994; Spencer *et al.*, 2003); to substantiate or illuminate the findings (Sandelowski and Barroso, 2002); or to allow readers to conceptualise possible alternative meanings and understandings (Elliot *et al.*, 1999). Several tools specified how data should be presented arguing that quotes should be clearly identified or numbered so that a reader can see that they don't just come from one or two people

[8] There were four exceptions to this. Beck (1993) also asked whether concepts, ideas and/or theory were well developed and whether readers find the findings meaningful and applicable in terms of their own experience. Blaxter (1996) also asked whether study findings were clear and distinguishable; whether findings addressed the aims of the study; and whether descriptions were detailed and context-rich. Elder and Miller (1995) also asked whether concepts, ideas and/or theory were well developed. Malterud (2001) also asked whether findings addressed the study aims.

(Blaxter, 1996; Boulton *et al.*, 1996; Britten *et al.*, 1995; Campbell *et al.*, 2003; CASP, 1998; Mays and Pope, 1995).

Just over half of the tools (n=18) included items related to the contribution of study findings to knowledge and/or the usefulness of findings for practice. Some tools contained items that simply asked whether the implications of findings for knowledge and/or practice had been considered/ adequately considered (Elliot *et al.*, 1999; Forchuck and Roberts, 1993; Long and Godfrey, 2004; Malterud, 2001; Sandelowski and Barroso, 2003). One tool linked usefulness to generalisability suggesting that if researchers had not discussed the transferability of findings it might be considered less valuable (CASP, 2002) and another tool linked the importance of findings in theoretical or practical terms to credibility (Blaxter, 1996).

With respect to whether study findings had made a useful contribution to knowledge some tools asked very general questions with very little or no supplementary guidance (e.g. has the study made a contribution to the discipline of family medicine) (Britten *et al.*, 1995; Cesario *et al.*, 2002; Elliot *et al.*, 1999; Mays and Pope, 2000; Vermeire *et al.*, 2003) whilst other tools asked appraisers to judge contribution to knowledge on the basis of whether or not previous knowledge and understanding are extended by the findings (Spencer *et al.*, 2003); new insights are offered (CASP, 1998; Malterud, 2001; Sandelowski and Barroso, 2002); or findings offer fertile ground for further research (Elder and Miller, 1995).

A similar pattern emerged with respect to items about the usefulness of findings for practice. Some tools asked fairly general questions with very little or no supplementary guidance on this aspect of study findings (e.g. are the findings relevant or important for practice?; is it clear what the implications for practice are?) (Britten *et al.*, 1995; Campbell *et al.*, 2003; Cesario *et al.*, 2002; Elder and Miller,

1995; Elliot *et al.*, 1999; Forchuck and Roberts, 1993; Kuzel and Engel, 2001; Long and Godfrey, 2004; Mays and Pope, 2000; Vermeire *et al.*, 2003). Other tools asked appraisers to judge usefulness to practice on the basis of whether or not study findings: helped them personally to, for example, understand their relationships with patients and their families (CASP, 1998; Elder and Miller, 1995; Giacomini and Cook, 2000a,b); specified a basis for action (Miles and Huberman, 1994); helped solve problems (Miles and Huberman, 1994); and/or helped to empower participants and equip them with new skills (Miles and Huberman, 1994).

Just under half of the tools included items asking about the scope and boundaries for making wider inferences from the study findings (n=14). Terms such as 'generalisability' and 'transferability' were sometimes used interchangeably across tools. For example, some tools favoured the term transferability (e.g. are the findings of this study transferable to a wider population) (Campbell *et al.*, 2003; CASP, 1998; Greenhalgh and Taylor, 1997) whilst others used the term generalisability (e.g. to what population are the study findings generalisable?) (Britten *et al.*, 1995; Kuzel and Engel, 2001; Long and Godfrey, 2004). One tool asked whether it was possible to assess the typicality of the study findings (Popay *et al.*, 1998) and two tools asked whether study findings, such as theory, hypotheses or propositions, would 'fit' contexts outside of the study situation (Beck, 1993; Whittemore *et al.*, 2001).

Several of the findings-orientated tools distinguished between different types of generalisability, or different types of wider inference (Elliot *et al.*, 1999; Popay *et al.*, 1998; Miles and Huberman, 1994; Spencer *et al.*, 2003). For example, Popay *et al.* (1998) distinguish between 'statistical' generalisability to a population and 'theoretical' or 'conceptual' generalisability to a similar class of phenomena, arguing that the latter is the goal of 'qualitative' research. Elliot *et al.* (1999) distinguish

between achieving a general understanding of a phenomenon, which they argue requires a range of informants and situations, and specific understanding of one case. Some tools highlighted the importance of 'thick description' (discussed below) for assessing generalisability (Elder and Miller, 1995; Miles and Huberman, 1994; Popay *et al.*, 1998; Spencer *et al.*, 2003).

As noted earlier, the methods-orientated tools tended to focus only on the three aspects of study findings, just discussed, which were assessed by the largest number of tools. The findings-orientated and methods- and findings-orientated tools went on to demand much more from the findings of studies than the methods-orientated tools. To start with, several of these tools wanted study authors to have provided clear statements of the findings of their study (Blaxter, 1996; Campbell *et al.*, 2003; CASP, 1998; CASP, 2002; Sandelowski and Barroso, 2002). Sandelowski and Barroso (2002, p 40), for example, required findings to be distinguishable from the data on which they were based and warned against "heaped description" whereby researchers present lots of quotes or case histories with no or minimal interpretation.

A concern with the quality of interpretation also featured in the 12 tools that included items to assess whether the analysis had fully developed concepts, ideas, or theory. The authors of the tools that included these items wanted to see concepts or themes which: were adequate (Popay *et al.*, 1998); precise and dense (Corbin and Strauss, 1990; Sandelowski and Barroso, 2002); connected with previous theory (Beck, 1993; Elder and Miller, 1995; Kuzel and Engel, 2001); could explain process and variation (Corbin and Strauss, 1990; Mays and Pope, 2000; Sandelowski and Barroso, 2002); were systematically related to each other (Cesario *et al.*, 2002; Corbin and Strauss, 1990; Elliot *et al.*, 1999; Miles and Huberman, 1994; Sandelowski and Barroso, 2002; Whittemore *et al.*, 2001); and described a

whole/comprehensive picture of the phenomenon under study included broader conditions and structures (Cesario *et al.*, 2002; Corbin and Strauss, 1990; Giancommini and Cook, 2000a,b).

The provision of a comprehensive picture of the phenomenon under study was at the heart of items in the nine tools that demanded detailed and context-rich descriptions. Several tool authors referred to this as 'thick description' (Miles and Huberman, 1994; Muecke, 1994; Popay *et al.*, 1998; Whittemore *et al.*, 2001). Some tools authors emphasised the purpose of such description as a way for the reader to gain a full understanding of the phenomenon or case under study without losing context (Blaxter, 1996; Cesario *et al.*, 2002; Elliot *et al.*, 1999; Giacomini and Cook, 2000a,b) or to portray richness and complexity (Spencer *et al.*, 2003). Others argued that thick description created the grounds for explanation (Popay *et al.*, 1998) and the understanding of difference in experience and perspective (Muecke, 1994).

For several tool authors 'thick description' and/or the full development of theory, with special attention to 'lay' or subjective meaning and diversity in perspective and experiences, should culminate in findings which illuminate participant perspectives and meanings in a believable way (Popay *et al.*, 1998; Spencer *et al.*, 2003; Whittemore *et al.*, 2001). For Whittemore *et al.* (2001) this was a sign that findings were 'authentic'. For others, key tests of the findings were whether findings resonated with readers' own experiences (Beck, 1993; Britten *et al.*, 1995; Cesario *et al.*, 2002; Elliot *et al.*, 1999; Miles and Huberman, 1994; Muecke, 1994; Giacomini and Cook, 2000a,b; Greenhalgh and Taylor, 1997); or whether participants found the findings to be an accurate, recognisable, honest and/or caring account of their experiences (Britten *et al.*, 1995; Miles and Huberman, 1994; Muecke, 1994).

e) Write-up and ethics

Individual tool items in this domain invited judgements on whether (number of tools containing items on this issue):

- The write-up is clear, coherent and easy to navigate (n=7).
- The write-up is comprehensive (n=3).
- Ethical issues have been considered and addressed/addressed adequately (n=11).

There were no differences in this domain between the different types of tools (methods-, findings- or methods- and findings-orientated tools). Tool items about the write-up were not concerned with how well methods or findings had been reported but with the quality and clarity of the manuscript (Cesario *et al.*, 2002; Elliot *et al.*, 1999) and whether or not it was "competent literature" (Muecke, 1994, p 197). Tools authors wanted to see good structure and signposting (Elliot *et al.*, 1999; Spencer *et al.*, 2003); technical terms defined (Elliot *et al.*, 1999); titles and headings that accurately reflect content (Malterud, 2001; Sandelowski and Barroso, 2002); and a literary style and form which fitted the methods, findings, and intended audience (Sandelowski and Barroso, 2002). Britten *et al.* (1995) argued that an unclear write-up might indicate unclear thinking and an underdeveloped analysis. Three of these tools also asked whether the write-up was comprehensive, covering all of the steps involved in the study (Cesario *et al.*, 2002), as well as a discussion of the limitations of the study (Malterud, 2001; Sandelowski and Barroso, 2002).

Some of the 11 tools that included items about ethical issues simply asked whether study authors had adequately addressed or considered ethical issues (Blaxter, 1996; CASP, 2002; Drisko, 1997; Miles and Huberman, 1994; Treloar *et al.*, 2000).

Other tools asked whether informed consent and ethical committee approval was obtained (Long and Godfrey, 2004; Cesario *et al.*, 2002); whether the researchers showed sensitivity and adequate respect for participants (Elliot *et al.*, 1999; Whittemore *et al.*, 2001); whether participants were informed of their rights and mechanisms were in place to protect these rights (Cesario *et al.*, 2002; Spencer *et al.*, 2003); whether benefits and risks were discussed with participants (Spencer *et al.*, 2003); whether confidentiality and anonymity were discussed with participants (Spencer *et al.*, 2003); whether benefits and risks of the study and recruitment and consent procedures were tailored to the specific needs of the reported study (Sandelowski and Barroso, 2002); whether all quotes used in the study report had analytical value and present participants fairly (Sandelowski and Barroso, 2002); whether the research serves the needs of the participants (Whittemore *et al.*, 2001); and whether appropriate information, advice, or service referral was offered to participants at the end of the study (Spencer *et al.*, 2003).

(v) Evaluation of tools

The comments from six systematic reviewers about their experiences of applying a sub-sample of the 31 tools to assess the quality of studies were grouped into tool strengths and weaknesses. Strengths and weaknesses were transformed into 10 desirable features of useful tools for use in systematic reviews and all 31 tools were judged according to whether these features were present or absent. None of the desirable features were present in all tools (range from 0 to 23 tools) and none of the tools demonstrated all of the desirable features (range from 0 to 7 desirable features) (table 5.3). Two tools stand out in table 5.3 as they met seven of the ten desirable features (Campbell *et al.*, 2003; Sandelowski and Barroso, 2002). These were both method- and findings-orientated tools designed for use in syntheses of 'qualitative' research. In addition both tools had undergone a period of development

Table 5.3: Evaluation of tools against ten features of useful tools identified by experienced systematic reviewers
(✓ = present; ✗ = absent)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1) Information provided to help reviewer judge the fit between the tool, the studies, and the appraisal context.	✗	✓	✓	✗	✓	✓	✓	✓	✗	✓	✗	✗	✓	✗	✓	✗
2) A distinction is made between the quality of the study and the quality of the report that describes it.	✓	✗	✓	✗	✗	✓	✓	✓	✗	✓	✗	✗	✗	✗	✓	✗
3) Guidance is provided to help reviewers arrive at the quality judgements required by the tool.	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	✗	✓	✓
4) A distinction is made between the collection of descriptive information and judging quality.	✓	✓	✓	✓	✗	✓	✓	✓	✗	✓	✓	✗	✓	✓	✓	✗
5) A means to record judgements about quality is provided.	✗	✗	✗	✓	✗	✓	✓	✓	✓	✗	✗	✗	✗	✗	✗	✗
6) Reviewers are helped to assess whether study conclusions are warranted given the study design, methods, and sample.	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
7) Evidence is provided for why particular characteristics of research are assumed to constitute 'good' or 'bad' quality.	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
8) Reviewers are helped to identify the 'fatal flaws' in a study.	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
9) Reviewers are helped to make an overall judgement about study quality.	✗	✗	✗	✗	✗	✓	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗
10) When an overall scoring system is provided it allows flexibility in the weight that can be given to individual items.	N/A	N/A	N/A	N/A	N/A	✓	N/A	N/A	✗	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Total number of features present in each tool	3	3	4	3	2	7	5	5	3	4	1	1	3	1	4	1

Key to tool numbers

1. Beck (1993)
2. Blaxter (1996)
3. Boulton and Fitzpatrick (1997)
4. Boulton *et al.* (1996)
5. Britten *et al.* (1995)
6. Campbell *et al.* (2003)
7. Critical Appraisal Skills Programme (1998)
8. Critical Appraisal Skills Programme (2002)
9. Cesario *et al.* (2002)
10. Corbin and Strauss (1990)
11. Drisko (1997)
12. Elder and Miller (1995)
13. Elliot *et al.* (1999)
14. Forchuck and Roberts (1993)
15. Giacomini and Cook (2000a,b)
16. Greenhalgh and Taylor (1997)

Table 5.3 (cont'd): Evaluation of tools against 10 desirable features of useful tools identified by systematic reviewers
(✓ = present; ✗ = absent)

	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	Total[a]
1) Information provided to help reviewer judge the fit between the tool, the studies, and the appraisal context.	✓	✗	✓	✓	✗	✗	✓	✗	✓	✓	✓	✓	✓	✓	✓	20
2) A distinction is made between the quality of the study and the quality of the report that describes it.	✓	✗	✗	✗	✗	✗	✓	✗	✗	✓	✓	✗	✗	✗	✓	12
3) Guidance is provided to help reviewers arrive at the quality judgements required by the tool.	✗	✗	✗	✗	✗	✗	✓	✗	✓	✓	✓	✓	✓	✓	✗	21
4) A distinction is made between the collection of descriptive information and judging quality.	✗	✗	✗	✗	✗	✗	✓	✓	✓	✓	✓	✓	✓	✓	✗	20
5) A means to record judgements about quality is provided.	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗	✓	✓	✗	✓	✗	9
6) Reviewers are helped to assess whether study conclusions are warranted given the study design, methods, and sample.	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗	1
7) Evidence is provided for why particular characteristics of research are assumed to constitute 'good' or 'bad' quality.	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	0
8) Reviewers are helped to identify the 'fatal flaws' in a study.	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗	1
9) Reviewers are helped to make an overall judgement about study quality.	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗	3
10) When an overall scoring system is provided it allows flexibility in the weight that can be given to individual items.	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	✓	N/A	2
Total number of features present in each tool	2	0	2	1	0	0	4	1	3	3	7	4	3	6	2	

[a]Total number of tools with feature present

Key to tool numbers

17. Hoddinott and Pill (1997)
18. Kuzel and Engel (2001)
19. Long and Godfrey (2004)
20. Malterud (2001)
21. Mays and Pope (1995)
22. Mays and Pope (2000)
23. McLaughlin (1986)
24. Miles and Huberman (1994)
25. Muecke (1994)
26. Popay *et al.* (1998)
27. Sandelowski and Barroso (2002)
28. Spencer *et al.* (2003)
29. Treloar *et al.* (2000)
30. Vermeire *et al.* (2002)
31. Whittemore *et al.* (2001)

and testing with feedback and advice from a wider group. (Campbell *et al.*, 2003 was based on the tools developed by CASP, 1998 and CASP, 2002 which each met five of the desirable criteria).

The first desirable feature listed in table 5.3 - that tools should provide information to allow reviewers to assess whether the tool fits with the studies to be appraised and the appraisal context – was present in nearly two-thirds of the tools. The seven tools that were designed for a particular type of 'qualitative' research could help reviewers to choose an appropriate tool on the basis of study type: evaluations (Long and Godfrey, 2004; Spencer *et al.*, 2003); ethnography (McLaughlin, 1986; Muecke, 1994); grounded theory (Corbin and Strauss, 1990); interviews (Hoddinott and Pill, 1997); and focus groups (Vermeire *et al.*, 2003). Reviewers could also make use of the definitions of 'qualitative' research that were provided by some of the tools to help them match tools to their appraisal context. Some tools gave very specific definitions of 'qualitative' research as studies that are designed to examine subjective meanings or the worldview of study participants (Blaxter, 1996; Boulton and Fitzpatrick, 1997; Britten *et al.*, 1995; Campbell *et al.*, 2003; CASP, 1998; CASP, 2002; Elliot *et al.*, 1999; Malterud, 2001; Popay *et al.*, 1998). These tools may not be useful for an appraisal context that requires the study of process and structure via, for example, participant observation. Other tools explicitly catered for a range of 'qualitative' research types, and tool items and guidance reflected this diversity (Giacomini and Cook, 2000a,b; Sandelowski and Barroso, 2002; Spencer *et al.*, 2003; Treloar *et al.*, 2000; Whittemore *et al.*, 2001). These tools can be contrasted with those that treated 'qualitative' research as a homogenous endeavour and as a consequence offered little useful information on whether their tool would fit the studies to be appraised or the appraisal context.

The second desirable feature listed in table 5.3 relates to the difficulty of disentangling the quality of studies with the quality of study reports. Systematic reviewers found it helpful when tools acknowledged this problem and made a distinction between the quality of the study and the quality of reporting. Across all 31 tools, only a third were judged to have made this distinction. Some tool items were designed specifically to assess the quality of study reports. For example, one item within the tool developed by Malterud (2001) asks reviewers to assess whether the title of the study report gives a clear account of the aim of the study. Other tool items, however, asked reviewers to assess the quality of the study and then directed them towards assessing the quality of the study report. For example item five in Vermeire *et al.* (2003) asks reviewers whether the sampling strategy was appropriate given the aims of the study and then directs reviewers to six sub-items all but one of which ask reviewers about reporting quality (e.g. has the relationship between subjects been described?). Other tool items confused the quality of study reports with the quality of the study within a single item. For example, the seventh item in the tool developed by Mays and Pope (1995) asks reviewers to judge whether the procedures for data analysis were clearly described [quality of reporting] *and* whether the procedures were theoretically justified [quality of study].

Reviewers found it helpful when tool authors had provided some guidance to help them to arrive at their quality assessments (the third desirable feature listed in table 5.3). Over two thirds of the tools provided this kind of guidance. The style of guidance varied across tools but was either structured or unstructured. Unstructured guidance took the form of a paragraph or two under the items in the tools (Boulton and Fitzpatrick, 1997; Britten *et al.*, 1995; Elder and Miller, 1995; Giacomini and Cook, 2000a,b; Greenhalgh and Taylor, 1997; McLaughlin, 1986; Popay *et al.* 1998; Treloar, 2000). These paragraphs covered a number of issues including: explanations for why tool authors considered that item to be important for assessing

quality; suggestions for what to look out for in the study report to help assess the study against an item; and statements about what the tool authors might expect to see in a high quality study/report. Structured guidance within tools helped reviewers to arrive at a judgement on each item or quality dimension covered by the tool by asking them to work through a set of detailed sub-questions or statements about what might be desirable/undesirable (Beck, 1993; Blaxter, 1996; Boulton *et al.*, 1996; Campbell *et al.*, 2003; CASP, 1998; CASP, 2002; Cesario *et al.* 2002; Corbin and Strauss, 1990; Elliot *et al.*, 1999; Muecke, 1994; Miles and Huberman, 1994; Sandelowski and Barosso, 2002; Spencer *et al.*, 2003; Vermeire *et al.*, 2003).

When tool authors provided guidance for reviewers they sometimes asked reviewers to collect descriptive information. For example, to help reviewers assess whether the study report included a clear statement of the study aims, CASP (1998) asks reviewers to collect information on what study authors were trying to find out. However, reviewers did not find it helpful if tool authors did not make a distinction between items that required the collection of descriptive information and items that required a quality judgement (the fourth desirable feature listed in table 5.3). Just over a third of the tools did not make this distinction. For example Long and Godfrey (2004) often listed descriptive items (e.g. within which geographical care setting is this study carried out?) alongside items to assess quality (e.g. is the setting appropriate with respect to the research question?) with no indication of how they related to each other or whether a reviewer should use the descriptive information to help assess quality.

As well as guidance on how to assess studies against each item covered by a tool, reviewers also wanted a means to record their judgements about quality (the fifth desirable feature listed in table 5.3). Only nine tools provided this and of these, five gave instructions on how reviewers should record their judgements. Boulton *et al.*

(1996) ask reviewers to make an assessment on a three point scale. For example, reviewers are instructed to answer 'yes', 'no', or 'uncertain' to the question 'are the criteria for selecting the sample clearly described?'. Campbell *et al.* (2003, p 674) instruct users of their tool to "answer yes/no if possible and then expand. Where appropriate, structure the expanded response by describing what the authors say, quoting directly where relevant (D) and then providing your own comments (C) on the clarity and aptness of the authors' description". The tools developed by CASP (1998) and CASP (2002) also provide yes/no options for reviewers to record their assessments. Sandelowski and Barroso (2002) ask readers to record whether the desirable attributes within each of their 'appraisal parameters' are present or absent. Reviewers wanted tools to help them assess the quality of studies in relation to the conclusions drawn by study authors (the sixth desirable feature listed in table 5.3). Reviewers felt that some of the tools they tried out had missed key problems with the design and execution of the studies in relation to the author's conclusions. For example, one reviewer questioned the claims made by one study author that their findings were generalisable to a wider population because their analysis was based on only a small number of cases (Aari, 2003). A second reviewer also questioned the claims made by another study author that their findings demonstrated a cause and effect relationship between an intervention and an outcome using a 'qualitative' research design (Felix-Ortiz *et al.*, 2000). None of the tools tackled this issue head on. As noted earlier in this chapter, just over half of the tools (n=18) contained a question about whether the study design and methods were appropriate to address the research question, but none of the tools asked whether the conclusions drawn were appropriate given the study design and methods. Similarly, although 18 tools asked whether the sampling strategy and/or the sample was appropriate given the research question, only one tool asked whether the conclusions were appropriate given the sample size and/or composition. This tool was from Sandelowski and

Barroso (2002) who asked whether the sample size could support the study findings.

None of the tool authors cited any empirical evidence for why the presence of absence of particular attributes are associated with 'good' or 'bad' quality in research (the seventh desirable tool feature listed in table 5.3). This is likely to be because very little empirical evidence exists. Greenhalgh and Taylor (1997) noted, for example, that they could not find any research on the effect of using more than one researcher to analyse data to justify the use of this strategy for increasing rigour. Although there was a lack of empirical evidence, some tools were better than others at providing explanations for why particular attributes made research better or worse (Corbin and Strauss, 1990; Elliot *et al.*, 1999; Giacomini and Cook, 2000a,b; Greenhalgh and Taylor, 1997; McLaughlin, 1986; Muecke, 1994; Sandelowski and Barroso, 2002). For example Giacomini and Cook (2000a,b) explained good and bad practice using a fictional research study, and Sandelowski and Barroso (2002) used examples to illustrate the consequences of failing to meet one of their appraisal parameters (e.g. "Interpretations of data are demonstrably plausible and/or sufficiently substantiated with data, as opposed to implausible as when a mother is quoted as hitting her child and this quote is used to illustrate the 'joys of motherhood'" (p40)).

Although all of the tools highlighted significant numbers of potential problems with studies, the tools did not help reviewers to identify what might be the 'fatal flaws' in studies or those problems which would cast serious doubt on the findings of studies (the eight desirable feature listed in table 5.3). There was, however, one exception to this. Sandelowski and Barroso (2002, p15) suggest that reviewers should try to distinguish between "non-significant representational errors and procedural interpretive mistakes fatal enough to discount the findings" and their tool asks

reviewers to indicate whether the presence of absence of the desirable features listed within their 81 appraisal parameters is relevant for assessing the quality of a particular study in their particular review context. The tool developed by McLaughlin (1986, p188) to apply to ethnography did highlight fatal flaws but only in one part of its tool related to data collection. This tool prescribed, rather than encouraged reviewers to think about, fatal flaws because it directed them to discard conclusions based on "weak data". Examples of 'weak data' given by McLaughlin (1986) included data collected early on at entry into the field; data based on information heard 'second hand'; data prompted by fieldworker question; or data gathered from respondents in the presence of others.

Only three of the tools required reviewers to make an overall judgement about the quality of a study (the penultimate desirable feature listed in table 5.3). Campbell *et al*. (2003) asked reviewers to report what their overall view was of the study and whether they would include it in their synthesis. At the end of the tool developed by Vermeire *et al*. (2003) two visual analogue scales are provided, anchored at each end by either 'yes' or 'no'. The first asks reviewers to rate whether the study is high quality and the second whether the study report should be published. Cesario *et al*. (2002) provided a scoring system within their tool to help reviewers arrive at an overall judgement. Their tool is organised around several categories, each containing several appraisal items requiring a yes/no response (e.g. Is essential descriptive information included?). Readers are instructed to score the study in each of the categories using a rating scale of 0 to 3, where 3 = Good = 75% to 100% of appraisal items met; 2 = Fair = 50% to 75% of appraisal items met; 1 = Poor = 25% to 49% of appraisal items met; 0 = No evidence that criteria met = <25% of appraisal items met. At the end of this process each study is assigned a 'level of evidence' on the basis of their total score across categories. However, reviewers

were not keen on this scoring system because it does not allow flexibility in the weight given to any one item in the tool (the last desirable feature listed in table 5.3).

5.4 Discussion

The survey of tools reported in this chapter located, compared and evaluated 31 tools for assessing the quality of 'qualitative' research. These tools appeared mainly in the healthcare literature and were devised largely by sociologists and/or academic doctors and nurses to help those unfamiliar with 'qualitative' research. Tools have been developed for the most part without funding and remain largely untested. There was considerable diversity and very little overlap amongst individual items within tools. It was not easy, however, to find an explanation for this diversity. There was no clear relationship between the type of items included in the tools and factors such as tool purpose, disciplinary background of tool authors, and the quality dimensions underpinning tools. It was difficult to characterise tools according to whether their content was underpinned by 'conventional' (e.g. reliability and validity) or 'alternative' quality dimensions (e.g. auditability, credibility). The same quality dimensions were defined in different ways by different tools and 'conventional' and 'alternative' quality dimensions were used interchangeably across and within tools.

When the items in tools were allocated and analysed within five study domains - background, theory and research questions; sampling, sample and setting; methods of data collection and analysis; findings; and ethics and write-up - three types of tool emerged. *Methods-orientated tools* had a predominant focus on fieldwork, data collection, and analysis and were often very specific with regard to strategies to increase rigour and reporting. *Findings-orientated tools* included less detail on methods but had more extensive coverage of findings. Whereas methods-orientated tools focused on whether findings were supported by data, generalisable and useful,

findings-orientated tools covered a greater range including: whether findings were clear and distinguishable, whether concepts or theory were well developed; whether diversity in meaning, perspective, and experience were captured; and whether findings resonated with readers and participants. *Methods- and findings-orientated tools* struck a balance between items about methods and findings, with some managing to retain a fairly detailed coverage of both domains.

These three types of tools, which emerged from an analysis of the actual content of tools, challenge the characterisations of tools to emerge from previous surveys or reviews of tools, which have emphasised the philosophical position of tool authors on the nature of knowledge and knowledge production in 'qualitative' research (Angen, 2000; Devers, 1999; Madill *et al.*, 2000; Murphy *et al.*, 1999; Spencer *et al.*, 2003). This study suggests that the philosophical explanation for differences in content across tools may have been overstated in previous work. In this study, tools differed according to the relative emphasis they placed on appraising quality through engaging with the methods used in the study or engaging with the study findings. Whether a tool was methods-orientated, findings-orientated, or methods- and findings-orientated did not relate in any straightforward way to the stated philosophical position of tool authors. This observation is similar to that made by Oakley (2000, p62) who also noted that tools with apparently different assumptions about how 'qualitative' research should be critically appraised reveal remarkable similarity in content.

The distinction to emerge in this study between quality assessment items which require reviewers to engage with study methods and those which require engagement with study findings has some resonance with a distinction made by Eakin and Mykhalovskiy (2003) between a 'proceduralist' and a 'substantive' approach to critical appraisal. A proceduralist approach, which they found to be

dominant in the 14 tools they surveyed, sees quality as arising from how well the research was carried out. In a substantive approach, however, there is a focus on analytical content as well as method so that, for example, an appraisal of data analysis would involve engagement with the findings of studies to understand how ideas, concepts or theory developed, as well as attention to the data analysis methods used. From this perspective, the methods-orientated tools of this study could be said to have adopted a proceduralist approach, whereas the methods- and findings-orientated tools would represent a more substantive approach. While Eakin and Mykhalovskiy (2003) clearly favour a substantive approach, this study made no direct attempt to evaluate the relative advantages and disadvantages of the three different types of tools. However, it is interesting to note that the two tools that displayed the most desirable attributes for practical application were methods- and findings-orientated tools (Campbell *et al.*, 2003; Sandelowski and Barroso, 2002).

Although two tools stood out from the others, in general the tools did not fare well with respect to how useful they might be for systematic reviews, or indeed how useful they might be in any context that requires their application to an actual study report. Common practical problems included a lack of provision for recording or making overall judgements on quality and confusion between study quality and reporting quality. More fundamental problems were no less rife. None of the tools provided any empirical evidence for the items they included and only one tool encouraged reviewers to either distinguish between minor and serious problems or check whether any conclusions drawn were warranted given the study methods and sample. Collectively, tools tended to 'sit on the fence' with respect to quality: reviewers would not be able to rely on any of these tools to help them make a decision on whether to include or exclude studies from a synthesis.

There are several possible reasons for why existing tools might not be very useful in application. The most obvious is the fact that very few tools were tested. Indeed the two tools that were judged to be the most useful both underwent quite extensive development, testing, and subsequent revision. Another reason is the lack of consensus amongst tools about how quality should be assessed. As each tool has been developed, the list of potential problems facing 'qualitative' research has grown and grown. Yet there appears to have been no effort to stem this proliferation by testing, for example, the various validation strategies suggested by tools or via thinking about how to distinguish between minor and major problems. This is despite the hopes of some of the earlier thinkers on quality in 'qualitative' research for such testing (Corbin and Strauss, 1990; Lincoln and Guba, 1985; Miles and Huberman, 1994). More recent thinking, however, has shown little support for the application of tools. There have been several high profile warnings about the 'dangers' of tools, with fears that they might stifle creativity and innovation in research (e.g. Barbour, 2001; Eakin and Mykhalovskiy, 2003; Torrance, 2004). Even tool authors themselves go to considerable lengths to highlight the disadvantages of using their tool (e.g. Elliot *et al.*, 1999; Spencer *et al.*, 2003). It is almost as if tool authors never actually intended their tool to be used to appraise the quality of studies at all.

Indeed, the findings of this study suggest that tools were in fact serving a purpose other than practical application. Their extensive coverage of potential problems, their attention to all stages of the research process, and their consideration of the multi-dimensional nature of quality, suggest that the tools are guidelines on designing, conducting and writing up research rather than aids to discerning the quality of a particular piece of research. As a large proportion of items in the tools were not specific to 'qualitative' research, the tools could in fact be read as guides to good scholarship in general rather than good 'qualitative' research. The tools can

also be characterised as 'manifestos' for 'qualitative' research, designed to educate and convince the reader of the value and merits of 'qualitative' research as compared to 'quantitative' research, or in some cases, the superiority of 'qualitative' research over 'quantitative' research.

The findings of this study need to be set against both strengths and limitations in the methods used to conduct the survey and evaluation. The strengths of this study lie in the exhaustive search strategy used to identify tools and the use of a systematic approach to collect and analyse data. These strengths contrast with previous work that has surveyed or reviewed tools in a selective way (e.g. Angen, 2000; Devers, 1999; Madill *et al.*, 2000). Another strength was the use of both 'qualitative' and 'quantitative' approaches to analysis, which combined analyses of definitions, meanings, and concepts within and across tools with analyses based on the frequency of occurrence of particular items or factors across and within tools. The use of either approach in isolation would have provided only part of the picture.

One possible limitation of the analysis of the content of the tools was the fact that only one researcher (myself) carried out the analysis. Although coding checks did identify errors introduced by boredom and fatigue (the sheer volume of data generated by the study at times overwhelmed me), engaging a second researcher would have made me more confident that the patterns I identified in the tools was an accurate representation of what was there. Another possible limitation of the study is that the list of desirable features of tools used in their evaluation was derived from a pool of researchers from the same institution. Although I tried to maximise as far as possible diversity of perspective within this group, the fact that all of these researchers worked on systematic reviews at the EPPI-Centre suggests a certain number of shared rather than diverse assumptions about the nature and purpose of research and systematic reviews. It may be that the desirable features of

tools identified in this study may only reflect the interests of researchers who hold similar views to those held by researchers at the EPPI-Centre. For example, it is unclear whether the same desirable tool features would emerge from discussion with researchers who reject outright the value of systematic review methodology for reviewing 'qualitative' research. Future work should consider repeating this part of the study with researchers who adopt a range of different perspectives.

A considerable amount of effort has already gone into the development of tools for assessing the quality of 'qualitative' studies. This survey has brought together and analysed this effort. One of the key messages from this study is that, despite being developed for practical application, existing tools for assessing the quality of 'qualitative' research are not well suited to this task. This is a surprising and unexpected finding which requires an explanation. One such explanation is that quality assessment tools are being used as a site for playing out the paradigm wars. Tools have not been designed to be applied at all but to demonstrate how 'qualitative' research is different from and/or better than quantitative research. One way to move beyond this position is to actually test the tools, and the quality criteria they cover, to build up a body of theoretical ideas and empirical evidence about what works best in which contexts. Including 'qualitative' research in systematic reviews offers an opportunity to do undertake such testing. The next two chapters describe work that makes use of such an opportunity.

CHAPTER 6

Study 2: An analysis of the development of a new tool to assess the quality of 'qualitative' research in systematic reviews

6.1 Introduction

Chapter five described a survey of tools to assess the quality of 'qualitative' studies. One of the surprising findings of this study was that many tools have been developed but few have been tried out. Detailed accounts of the development of tools are patchy in the reports describing them and tool authors rarely offered critical reflections on their tools. This lack of formal evaluation, detailed description and reflection makes it difficult to draw out lessons for further development. This chapter reports an analysis of the development of a new tool for assessing the quality of 'qualitative' studies. This tool helps reviewers to: a) collect relevant methodological and contextual information from study reports (e.g. why and how the study was carried out); b) judge whether a study meets a number of generic quality criteria (applicable to any type of study) and specific quality criteria (applicable to particular study types and research questions); and c) make an assessment about how much weight should be given to a study with respect to whether it is likely to provide a trustworthy answer to the review question under study.

A team of researchers based at the EPPI-Centre developed the tool. It is embedded within a broader framework for conducting systematic reviews that include diverse study types, and was developed across a series of systematic reviews in health promotion and public health (HP&PH) and a parallel programme of work to develop systematic reviews in education. (These two programmes of work were described in

chapter three.) The HP&PH reviews included, alongside trials, studies examining intervention processes or people's perspectives and experiences of particular health issues. The education reviews included a broader range of study types categorised according to whether they were 'descriptive'; an 'exploration of relationships'; a 'researcher-manipulated evaluation'; or a 'naturally occurring evaluation'.

The tool, and the systematic reviews of diverse study types in which the tool was used, are examples of methodological innovation in social science research. The first version of the tool was developed and applied in a systematic review in 1998, a time when there was interest in including 'qualitative' research in systematic reviews but few practical examples (Popay *et al.*, 1998; Speller *et al.*, 1997). The emphasis in the later versions of the tool on assessing quality in relation to research questions is also unique in the literature today. Previous tools to assess the quality of 'qualitative' research have started from the method (i.e. how can we assess the quality of 'qualitative' research?) or epistemological position, whereby 'qualitative' research is seen as a totally different enterprise to quantitative research (i.e. how can we show how it is different/better than quantitative research?). The analysis reported in this chapter aimed to address two questions: i) how did this methodological development happen?; and ii) what are the lessons to be learnt for fostering methodological development in the future?

6.2 Methods

As one of the tool developers I used my own knowledge to build a timeline charting the stages in the development of the approach. I identified and then reconstructed the stages in detail by drawing on several data sources:

- paper and electronic versions of the tool at different stages of development;

- the systematic review reports the tools were used in;
- peer referee comments on the systematic reviews;
- minutes of meetings;
- correspondence with funders; and
- correspondence between team members.

The timeline and detailed description of the development were used as basis for the analysis to identify factors driving or supporting the methodological development. Some of these factors began to emerge very early on in writing the detailed reconstruction of events. For example, in order to create a meaningful account it was necessary to describe the people involved in the team – their disciplinary backgrounds, intellectual interests and previous work. It quickly became apparent that the fact that the team was a multi-disciplinary one was a significant factor for the methodological development. I drew up a preliminary list of factors which were subsequently revised and expanded through collective reflection with two of the other key figures in the team. These two figures were Ann Oakley and Sandy Oliver, the supervisors of this PhD. I asked Ann and Sandy to go beyond their role as my supervisors to become 'key informants' for the analysis because their perspective was important for understanding the role of the early history of SSRU and the EPPI-Centre in the methodological development. (Ann is the founding director of SSRU and Sandy joined SSRU in 1993). However, the work presented in this chapter is still the product of my own original work. I remained responsible for the design, conduct, interpretation and write-up of the analysis at all times.

6.3 Results

(i) Overview

The new tool went through a number of stages of development over a period of eight years (figure 6.1). The associated review work is noted below the timeline whilst significant milestones in the development of the tool are noted above. These milestones were a mixture of key decisions; the construction and revision of guidelines; and associated developments from the EPPI-Centre systematic review work in education.

There was no specific funding to develop the tool. The HP&PH reviews shown below the timeline in Figure 6.1 were funded by the English Department of Health (DH) to provide timely answers to substantive questions about health promotion rather than methodological questions. The development was therefore driven by the need to address the substantive questions of interest rather than the pursuit of methodological innovation *per se*. Because there was no funding for development work, it had to be done fast within tight time-scales to meet the specific purposes of each review. The team working on the HP&PH reviews varied in size and composition at different stages of development. Whilst the team generally grew in size across time, some team members moved on to pastures new whilst others moved onto other projects temporarily (for example, collectively the team had four babies over the time period in question).

A number of factors were identified as influencing the development of the tool (figure 6.2). These factors were inter-related and played out in various ways at different stages of the timeline. The factors, and the events in the timeline, are described in more detail below.

Figure 6.1: Timeline depicting the development of the EPPI-Centre approach to assessing the quality of 'qualitative' and other types of studies

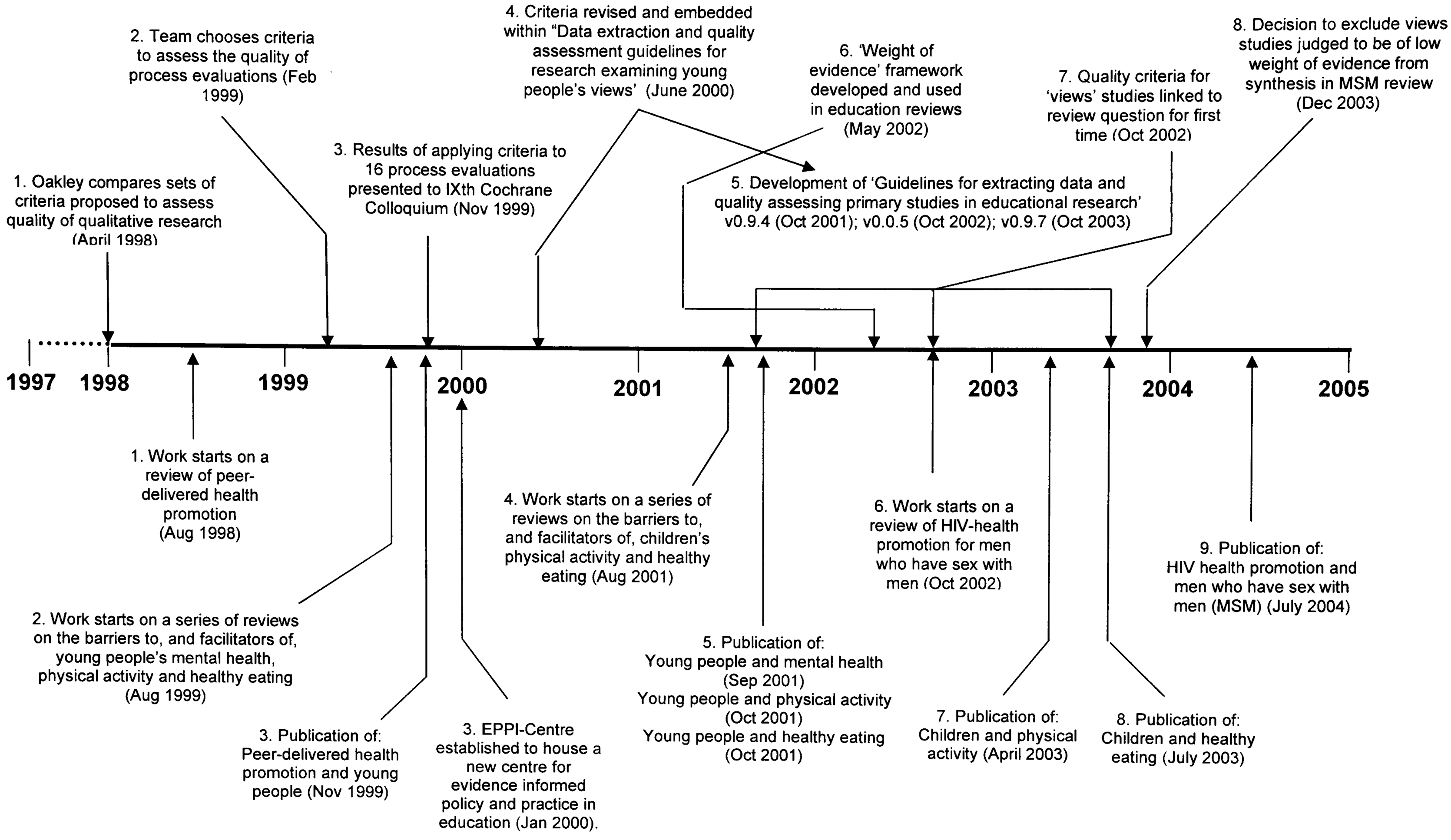


Figure 6.2: Factors influencing the methodological innovation behind the development of a new tool to assess the quality of 'qualitative' and other types of research in systematic reviews

- **Policy and funding climate**
  - Working with real questions
  - Open dialogue and negotiation between funders and research team
  - Product differentiation in a competitive market
- **Shared principles and interests amongst team members**
  - A desire to do methodological development
  - Production of relevant research (for policy, practice and personal decisions)
  - Working beyond the 'paradigm wars' – a commitment to RCTs *and* 'qualitative' research
- **Multi-disciplinary team**
  - Lateral thinking between disciplines
  - Challenge discipline blind spots
  - Stepping outside disciplinary boundaries
- **Quality of research**
  - The poor methodological quality of much of the 'qualitative' research in health promotion and public health
- **Ways of working**
  - Balance between individuals and team
  - Fast but creative thinking to 'get the job done'
  - Non-hierarchical structure

(ii) The early years

The timeline in Figure 6.1 starts in 1998 with a review of peer-delivered health promotion that included 'qualitative' and other types of studies of intervention processes alongside trials of intervention effects. This review was one of the key 'moments' in the development of the approach to assessing the quality of 'qualitative' and other types of studies that is described in this chapter. However, the work of SSRU and the EPPI-Centre prior to this review is just as important for understanding why and how this approach developed as it did. Established in 1990, SSRU has a history of conducting policy relevant research with a focus on health, education, welfare and other services, and the relationships between professionals

who deliver these services and the public who use them. By the time the review of peer-delivered health promotion had been commissioned, SSRU and the EPPI-Centre had already developed expertise and tools for conducting systematic reviews of social interventions. Ann Oakley (the founding director of SSRU and the EPPI-Centre) and colleagues have described how their work to develop systematic review methods for social interventions started with a small project funded by the UK's Economic and Social Research Council (ESRC) in 1993 to "provide for the social science research community a resource of evidence on the effectiveness of intervention and future research needs" (Oakley *et al.*, 2005, p 9). Early work included a project to develop a database of controlled trials in education and social welfare funded by the Economic and Social Research Council (ESRC) and reviews on the effects of anti-smoking, reading recovery and juvenile delinquency programmes. A series of grants to conduct systematic reviews in health promotion continued this early work and by 1997 a sustainable system for coding, classifying and appraising research relevant to the evaluation of health promotion and other types of social interventions had been established.

Specific components included: a keywording strategy to code research papers (Peersman and Oliver, 1997); a set of '*Review Guidelines*' to assess the quality and extract data on the findings of evaluation study reports (Peersman *et al.*, 1997); and specialist reviewing software and databases which have become known in more recent years as EPPI-Reviewer and EPIC respectively (Thomas, 2002). Another 'product' was a growing team of (contract) researchers with expertise in the conduct of systematic reviews, a capacity that was (and still is) rare in the social science community in the UK. As noted earlier in this thesis, this is illustrated by the fact that the EPPI-Centre is the 'Methods for Research Synthesis' node of the ESRC National Centre for Research Methods, which aims to develop methods and increase skills and capacity amongst social science researchers in the UK.

The above developments were guided by three important methodological principles: a question-led approach to conducting research which starts from the research question rather than methodology; ensuring that research questions are relevant to research users; and the appropriate use of both 'qualitative' and 'quantitative' methods in the systematic review and primary evaluation of interventions. These two principles are reflected in the '*Review Guidelines: Data collection for the EPIC database*' developed by Greet Peersman, Sandy Oliver and Ann Oakley (Peersman *et al.*, 1997) (Appendix A). Peersman and Oliver joined Oakley in 1995; Oliver via her work on user involvement in systematic reviews within the Cochrane Collaboration, and Peersman via her work in public health and research in various settings including the World Health Organisation (WHO)[1]. Both Peersman and Oliver had degrees in the biological sciences and had worked as practitioners to promote health. Peersman had worked in health education in Zimbabwe and Oliver had worked as an antenatal teacher for the National Childbirth Trust in the UK. For Oliver, this experience had been the catalyst for her interest in developing ways to enable service users to influence what research is undertaken and how.

At first glance Peersman's and Oliver's backgrounds would appear to be in stark contrast to Oakley's background. Oakley, a feminist sociologist, had a first degree in Politics, Philosophy and Economics and a PhD in Sociology. However, unlike many of her peers in sociology she had challenged the 'paradigm wars' that encouraged feminist and other researchers to use 'qualitative' methods over 'quantitative' ones. By the early 1990s Oakley had carried out numerous research projects, including small-scale 'qualitative' projects as well as large-scale RCTs. Despite their different disciplinary backgrounds, Oakley, Oliver and Peersman all shared a set of

---

[1] Together with James Thomas, who joined SSRU in 1994, Ann, Sandy and Greet established the Centre for the Evaluation of Health Promotion and Social Interventions (the EPI-Centre) in 1995. (The EPI-Centre became the EPPI-Centre in 2000.)

assumptions about the scientific method and the role of research. They believed it to be important to minimise bias and error in research and to ensure that research was relevant, accessible, and conducted with the involvement of the people whom the research was intended to benefit. Shared assumptions, despite different disciplinary backgrounds, continued to be characteristic of the team as it grew. Disciplines covered included sociology, social policy, statistics, information science, psychology, economics, biology, geography, education, and nursing.

The Review Guidelines were one of the first products Oakley, Oliver and Peersman worked on together. (The first version of the Review Guidelines were developed in 1993 by Ann Oakley and Deidre Fullerton.) The guidelines can be applied to three categories of 'intervention studies'[2]: 1) those which describe an intervention and its development; 2) those which also evaluate the feasibility, acceptability and implementation of the intervention (process evaluations); and 3) those which also evaluate an intervention according to whether it produces changes in specified outcomes (outcome evaluations). The guidelines encourage the collection of information from evaluation studies in a standardised and systematic way to minimise error and bias. Users of the guidelines are provided with a series of questions to answer about an evaluation study. These questions are organised into seven sections: study identification; support for the study; description of the intervention; description of the study population; planning and process measures; and quality of the outcome evaluation. In the final section of the guidelines the reviewers collect methodological information from an outcome evaluation report (e.g. How were participants allocated to intervention and comparison groups?; What outcomes did the authors say they were intending to measure?). Reviewers are

---

[2] Intervention studies are defined in the guidelines as those in which "the researcher attempts to change people's experience or situations by, for example, exposing people to an education programme, a skills training, or the use of a particular service, or by carrying out environmental modifications" (Peersman *et al.*, 1997, p 2)

then asked to make a judgement about whether the evaluation is 'sound' or 'not sound' on the basis of four criteria: (an equivalent control or comparison group; pre-intervention data reported for all individuals/groups; post-intervention data reported for all individuals/groups; and all outcomes reported on). When relevant, reviewers are prompted to consider the information they collected about the evaluation earlier. The guidelines help the reviewer be as transparent as possible in their judgements about quality. This is important because in systematic reviews studies are usually included or excluded into a synthesis on the basis of these judgements about quality.

The planning and process section of the Review Guidelines is novel in comparison to other systematic review tools for data extraction and quality appraisal. Eleven questions on planning are included to enable reviewers to collect data about who was involved in the development of the intervention (e.g. whether the intervention was based on a needs assessment, and if so what kind; who was involved in setting the aims of the intervention) and who was involved in the development of its evaluations (e.g. who identified the range of outcomes/processes to be addressed; how were the evaluators selected). The inclusion of these questions codifies the extent of user involvement in evaluation studies and makes this neglected aspect of research more visible. Other available systematic review tools for data extraction and/or quality assessment do not facilitate this with their focus on issues of research design and methods of data collection and analysis (see, for example, Cook *et al.*, 1995; the What Works Clearing House Evidence Standards[3]).

A further seven questions are included for extracting data on the methods and findings of the process evaluation. In contrast to outcome evaluations, the process

[3] Retrieved 15th January 2006 from http://www.whatworks.ed.gov/reviewprocess/standards.html

evaluation section does not include a structured format for helping reviewers to assess their quality. Two questions do, however, encourage the reviewer to think about quality: 'Are there any obvious inconsistencies in the reporting of the evaluation data' and 'Do the data presented substantiate the author's findings?' An important contribution of this section lay in its careful unpacking of the different types of questions that process evaluations could answer (in contrast to the single type of questions answered by an outcome evaluation). Nine types of processes are described in the guidelines and reviewers are asked to indicate which one(s) a particular evaluation addresses: acceptability of the intervention; accessibility of the intervention; consultation, collaboration and partnerships; content of the intervention; implementation of the intervention; costs associated with the intervention; management and responsibility; quality of the programme materials; and skills and training of the intervention providers.

Earlier versions of the Review Guidelines did not include a section on planning and process measures. The impetus to develop such a section came from the team's recent experiences in conducting and disseminating systematic reviews in sexual health promotion (Oakley *et al.* 1995, 1996; Peersman *et al.*, 1996) and in running critical appraisal workshops for health promotion practitioners called 'Promoting Health After Sifting the Evidence' (PHASE)(Oliver *et al.*, 1996). When conducting the reviews, it became apparent that the way interventions had been developed and implemented varied enormously and this variation appeared to be related to effectiveness. To understand this better, the team realised that they needed to collect information on development and implementation in a systematic way. At the same time, those in the health promotion field were arguing that systematic reviews should address development and implementation issues as well as effectiveness (e.g. Speller *et al.*, 1997; Peersman *et al.*, 1999). Oliver had an opportunity to develop work on these issues when designing materials for the PHASE workshops.

The team knew that tools developed to facilitate evidence-based medicine would not be acceptable to those working in the field of health promotion because they focused on RCTs and effectiveness. Oliver therefore developed a series of critical appraisal tools for the workshops including one for process evaluations.

One outcome from the workshop was the revision of an existing Cochrane review on the effects of smoking cessation programmes for pregnant women. Participants at the PHASE workshops had been very critical of this review; they argued that its focus on birth weight ignored other relevant outcomes such as mothers' stress and coping, and that no information was provided about the practicalities of developing and implementing smoking cessation interventions. Oliver suggested that they write to the author of the review with suggestions for how it could be improved. The author responded positively and invited Oliver to revise the review with her in order to incorporate women's views about the outcomes, as well as the processes of the smoking cessation interventions (Oliver, 2001). Meanwhile, Peersman had been busy building on the process evaluation tool constructed for the PHASE workshops to create the planning and process section of the guidelines. The Review Guidelines were eventually published in 1997 and they continue to be used in the same form in EPPI-Centre reviews in HP&PH today, ten years later.

(iii) Developments with process evaluations in health promotion

An opportunity arose in August 1998 to develop the work on process evaluations further when EPPI-Centre health promotion staff were asked to draw up a proposal for a systematic review on peer-delivered health promotion for young people as part of a three year programme of systematic review work funded by the DH from 1998 to 2001. This review had been chosen from a 'menu' of choices for policy makers of six different systematic reviews of health promotion. Nearly all of these review

suggestions offered the DH a product which went beyond a straightforward effectiveness review to include the 'non-trial' literature. There were several factors influencing this. As noted above, our previous experience suggested that policy makers and practitioners had questions which could not be answered by trials. They were also disappointed with 'empty' reviews which had not found much or any good quality evidence. Within academic circles too, questions were being raised about the role of 'qualitative' research in systematic reviews and evidence-based health promotion. Another important factor was our interest in doing methodological work. Including 'qualitative' research in systematic reviews presented intellectual as well as practical challenges and we were very keen to contribute solutions to these challenges. Proposing different review 'products' had also fulfilled the need to demonstrate to funders that what we planned to do was different to, or did not overlap with, the review work of other organisations. In other words we had been engaged in 'product differentiation' in a 'competitive market'.

We had suggested a review of peer-delivered health promotion as one of six possible review topics in our proposal to the DH because at this time no such review existed. This was a significant gap as the method was becoming hugely popular and was being widely used in numerous health promotion projects around the UK, Europe and worldwide and great claims were being made regarding both its efficacy and its potential to actively engage young people in promoting their own and others health (Milburn 1995; Svenson 1998; Wilton *et al.* 1995;). Our earlier reviews on sexual health also highlighted a peer approach (Oakley *et al.*, 1995; Peersman *et al.*, 1996) and led to the 'Ripple Study', a large randomised controlled trial of peer-led sex education in secondary schools in England (Stephenson *et al.*, 2004). We knew the DH was interested in peer-led approaches for health promotion as they had funded the pilot study at SSRU for the main 'Ripple Study' (Charleston *et al.*, 1996). A DH representative was also part of the steering group for the 'Ripple

Study'. (This DH representative was the same person who managed the research contract for our DH funded programme of systematic review in HP&PH.) Policy-makers at the DH wanted to know whether this approach worked, but also wanted to know about other related issues such as implementation and acceptability, and we were asked to draw up a proposal for the review.

By this time the team had grown to include two new researchers: myself and Ros Weston. I had a first degree in psychology, which had given me a good grounding in experimental design and statistics, and an MSc in psychology and health. The latter had offered me an opportunity to pursue theories and methods from sociology and philosophy as well as psychology. This opportunity also meant exposure to the 'paradigm wars' especially with the recent emergence of the 'critical psychology' movement (e.g. Parker, 1989). Although I pursued some of these ideas in a study examining how young people talked about contraception (Harden and Willig, 1998), I quickly became disillusioned with arguments against 'quantitative' methods, realising that I felt more comfortable using diverse methods depending on the research problem in hand. My first experience of conducting funded research for a regional health authority (on the topic of young people and sexual health) stimulated interest in conducting research for policy and practice and listening to the views and experiences of the public (e.g. Harden and Ogden, 1999a,b,c). Although I had never heard of evidence-based health promotion or systematic reviews, the EPPI-Centre at SSRU appeared to provide the perfect environment for me to pursue these interests further when I applied for a research officer position at SSRU in 1997.

Weston originally trained as a teacher before she became a Lecturer in Health Promotion within the School of Education at the University of Southampton. Before joining the team at the EPPI-Centre in 1998 she had already been conducting systematic reviews, collaborating on a Cochrane review of cervical cancer

prevention using EPPI-Centre methods with Peersman and Jonathan Shepherd (Shepherd *et al.*, 2000). (Jonathan Shepherd later joined the EPPI-Centre team on a secondment basis from the Wessex Institute for Health Research and Development at the University of Southampton. His background and role is discussed in more detail later.) Weston was part of an established network within the field of health promotion having previously worked at the Health Education Authority for England and as the director of a Europe Against Cancer funded research project to examine ways forward for the evaluation of cancer prevention interventions. She used the latter project as a case study for her PhD on evidence-based health promotion in which she argued that an evaluator's preferred method of evaluation is often linked to his/her original disciplinary training and/or preferred model of health promotion (e.g. 'the prevention model'; the 'educational model'), rather than being chosen as the most appropriate method for the evaluative question in hand (see Weston, 1998). Oakley had also been writing on a similar theme exploring reasons for resistance to experimental evaluation from those working in health promotion (Oakley, 1998).

We drew up a proposal for the review of peer-delivered health promotion and sent it to the DH at the end of August 1998. The proposal outlined our intention to determine a) the effectiveness of peer-delivered health promotion using well-designed outcome evaluations and b) the appropriateness of this type of intervention using process evaluations, formative evaluations, needs assessments and descriptions of the development of interventions (with a note explaining that these types of studies are "likely to include large amounts of qualitative data"). The proposal described how appropriateness would be examined in terms of whether the self-defined health promotion needs of young people are fulfilled by this type of intervention, its acceptability and accessibility, and the barriers and facilitators to intervention implementation. Standard stages of a systematic review were outlined

including exhaustive searching and quality assessment of studies. Established criteria were offered for the critical appraisal of outcome evaluations and the proposal outlined our intention to synthesise the findings on appropriateness within a "quality control framework informed by current work in progress on guidelines for the assessment of qualitative data".

Consultation with external groups was a key part of the review process. The DH sent out the proposal for peer referee and three sets of comments were received in early November. Our intention to include process evaluations was welcomed by all three referees with one suggesting that their inclusion would make the review more relevant to practitioners (with the comment that review users need to know 'how to do it' as well as 'does it work'). As a consequence of including process evaluations, peer referees also highlighted the need for methodological development to take place during the review; one referee was unclear about how we were going to critically appraise the process evaluations and another suggested that "the development of protocols [for process evaluations] .....is interesting and should be open to debate during the course of the study". Procedures for ensuring that our protocols were open to debate were already in place; throughout the review process we consulted with potential users of the review on its methodology and scope via an established steering group for the wider three-year programme of work. This group, which included policy makers, service providers and other researchers, met twice during the review period. The steering group acknowledged the importance of the development of quality criteria for process evaluations and thought that this work should be a necessary part of the review.

By December 1998 exhaustive searches had been completed, the resulting citations had been screened for relevance, and descriptive coding had been done on those studies that met the inclusion criteria for the descriptive mapping stage of the

review. A total of 77 stand-alone process evaluations had been found (with a further 75 embedded within outcome evaluations). One of the team (Ros Weston) had begun to think about how we might a) deal with such a large number of diverse studies and b) assess their quality. Weston had recently completed a selective review of evaluation studies of cancer prevention interventions. This review included diverse types of evaluations, some of which used 'qualitative' methods and Weston had developed six criteria to select studies based on the literature on evaluation methods for health promotion (e.g. Nutbeam, 1999; Tones and Tilford, 1994). These criteria required an evaluation to: 1) include an explicit account of the theoretical underpinning/ generic framework for the intervention and evaluation; 2) include a formative evaluation (needs assessment phase and pilot); 3) include an intermediate evaluation (a review at the midpoint of the intervention); 4) include an impact (short-term) evaluation; 5) include an outcome (long-term) evaluation; and 6) add value to existing knowledge and practice.

Weston recommended that we apply a modified list of these criteria as an initial methodological screen for the process evaluations identified for the review of peer-delivered health promotion. Focusing their efforts on the 77 stand alone process evaluations, two of us (myself and Weston) modified and agreed the criteria and then assessed each of the studies according to whether it was a 'formal' process evaluation[4]. In practice this meant that the studies had report a formative evaluation (conducted prior to the implementation of an intervention); an intermediate evaluation (conducted at the mid-point of an intervention); or a summative evaluation (conducted at the end of the intervention) AND the study report had to

---

[4] We focused on process-only evaluations because we suspected that there would be few good quality outcome evaluations with integral process evaluations. This was indeed the case as only four of the 12 good quality outcome evaluations had integral process evaluations, reflecting a more general trend in the UK and elsewhere to carry out process only or outcome only evaluations. This scenario mirrors traditional distinctions between 'qualitative' and 'quantitative' researchers.

include a clear and systematic presentation of the evaluation methodology and results. These criteria proved to be extremely useful for making sense of the diverse set of literature coded as 'process evaluations'. Using them we found that 23 studies which had been initially classified as process evaluations could not be considered a 'formal evaluation'. (Thirty-nine process evaluations had already been excluded, as they were not within the scope of the in-depth review.) Although the reports of these 23 studies did suggest that an evaluation had been carried out, the methods and results were not clearly or systematically presented. For example in one unpublished report initially classified as a process evaluation (McGuiness, 1994, p6), a peer-led intervention was described and a few evaluative comments were made along the way in the study report (e.g. "After each session club members were asked to fill in a questionnaire. Results show that they found the sessions fun, relevant and useful with an overwhelming 80% wanting the team to call again").

Sixteen process evaluations passed this initial methodological filter but we still needed to decide how to assess their quality in more detail. We presented some preliminary thoughts on quality criteria to the first meeting of our steering group in December 1998, suggesting that we look at existing literature on assessing the quality of 'qualitative' research. We approached this literature from a particular viewpoint. Although the team had considerable experience of 'qualitative' research, we did not identify ourselves as 'qualitative researchers' and, as already noted, were critical of the paradigm wars. Around this time, Ann Oakley was close to a draft of her book '*Experiments in Knowing*' about the history and sociology of research methods in the social sciences (Oakley, 2000). A key aim of this book was to illustrate how decisions about research methods are influenced by much more than the technical issue of which method is best able to answer the research question in hand. Integral to this was the exposure of the ideological basis of the

'paradigm wars' in which 'quantitative' ways of knowing are often rejected in favour of 'qualitative' ones. Oakley had observed this happening within health promotion where many evaluators and practitioners favoured 'qualitative' evaluations regardless of the evaluative question to be addressed (Oakley, 1998). In the third chapter of the book Oakley examined the problem of how we might distinguish between good and bad 'qualitative' research. We read the chapter to help inform decisions about how to assess the methodological quality of process evaluations.

Oakley had examined eight sets of criteria divided into two groups: 1) those sets of criteria which take the position that 'qualitative' research can be judged by the same principles as 'quantitative' research (Blaxter, 1996; Boulton *et al.*, 1996; Cobb and Hagemaster, 1987; Mays and Pope, 1995); and 2) those sets of criteria which assume that 'qualitative' research is an altogether different form of inquiry and, accordingly, should be judged by different criteria (Lincoln and Guba, 1985; Leininger, 1994; Muecke, 1994; Popay *et al.*, 1998). We adopted the first position that 'qualitative' research is not a completely different enterprise and therefore should be judged according to the standards proposed for any type of scientific research. We examined the four sets of criteria in the first group to identify possible candidates to use for quality assessing the process evaluations (table 6.1).

The seven criteria listed in the first column of table 6.1 were selected. Making this selection was not an easy task. We could not use those criteria that were most commonly advocated across the four existing tools as a basis for selection. As Oakley (2000, p 57) had already noted, there was little overlap in the tools: in total they advocated 46 <u>different</u> criteria, only two of which were common to all four sets.

Table 6.1: Origins of criteria used to assess the quality of process evaluations in a review of peer-delivered health promotion

Process evaluation tool	**Cobb and Hagemaster (1997)**	**Boulton et al. (1996)**	**Mays and Pope (1995)**	**Blaxter (1996)**
Background to study/theoretical framework				
1. An explicit theoretical framework and/or literature review.	*Understanding of qualitative paradigm *Appropriate references cited *Inclusion of a literature review *Literature review sufficiently comprehensive *Appropriate initial framework *Knowledge of qualitative research strategies demonstrated *Understanding of ethical issues * Importance of study to subject area outlined	*Study located in broader context	*Explicit account of theoretical framework and methods stated	*Clear connection to existing body of knowledge
Aims/research questions				
2. Aims and objectives clearly stated.	*Problem clearly defined *Study purposes clearly stated *Major concepts defined *Purpose of research is discovery/description/theory building/illustration *Scope of question manageable within study time frame	*Clear aims *Qualitative approach appropriate	No criteria specified	*Research methods appropriate to research questions *Sensitivity of methods matches needs of research questions
Study context (including role of researcher)				
3. A clear description of context.	*Context adequately described *Researchers-respondent relationship understood *Role of researcher apparent		*Clear description of context	*Relationship between researchers and subjects considered and research explained to subjects *Clear contextualisation of research *Clear statement of author's own position
Sampling methods and sample				
4. A clear description of the sample and how it was recruited.	*Characteristics of sample outlined *Issues of qualitative study sampling adequately addressed	*Clear description of sample *Clear description of recruitment *Adequate and appropriate final sample	*Clear description and justification of sampling strategy *Theoretically comprehensive sampling strategy to ensure generalisation of conceptual analyses	*Clear criteria for sample selection and data collection and analysis *Theoretical justification for selection of cases

Table 6.1 (cont'd): Origins of criteria used to assess the quality of process evaluations in a review of peer-delivered health promotion

Process evaluation tool	Cobb and Hagemaster (1997)	Boulton *et al.* (1996)	Mays and Pope (1995)	Blaxter (1996)
Data collection and data analysis methods				
5. A clear description of methods used to collect and analyse data.	*Plan for gaining entrée given *Plan for organising/ retrieving data outlined *Framework for analysis stated	*Adequate description of fieldwork *Adequate description of data collection methods *Adequate description of data analysis	*Clear description of fieldwork methods *Clear description and theoretical justification of data analysis procedures	*Systematic data collection and record keeping.
Ensuring reliability/validity/trustworthiness of data analysis				
6. Analysis of data by more than one researcher.	*Problems of reliability and validity addressed	*Evidence that supporting material is representative *Evidence of efforts to establish validity *Evidence of efforts to establish reliability	*Analysis repeated by more than one researcher *Use of quantitative evidence to test qualitative conclusions where appropriate *Evidence of seeking out contradictory observations	*Systematic analysis *Adequate discussion of evidence for an against researchers' arguments *Measures taken to test validity of findings *Steps taken to see if analysis is comprehensible to participants *Systematic presentation of data *Reference to accepted analytic procedure
Tracing path between data interpretations and conclusions				
7. Inclusion of sufficient original data to mediate between evidence and interpretation.	*Demonstration of how framework is derived from data	*Evidence provided to support analysis *Sufficient original material presented	*Independent inspection of evidence possible *Sufficient original evidence presented to satisfy reader of relation between interpretation and evidence	*Adequate discussion of how findings are derived from the data *Clear distinction made between data and interpretation *Sufficient original evidence presented to satisfy reader of relationship between evidence and conclusions *Credible and appropriate results

Listing the seven criteria which were selected to assess process evaluations according to the stage of the research process they referred to, and re-organising the criteria advocated in the four sets of criteria on which these seven were based, illustrates the relationship between the different tools and the rationale for the selection.

For 'background to the study and theoretical frameworks'; 'study context', 'data collection and data analysis methods', and 'tracing a path between data, interpretations and conclusions', requests for 'clear', 'adequate', 'sufficient' or 'explicit' descriptions or statements were virtually the only desirable characteristics demanded across the sets of criteria (see rows 1, 3, 4, and 7 in Table 1). For 'ensuring the reliability and validity of data analysis' (row 6) 'analysis of data by more than one researcher' was chosen in favour of any of the other criteria advocated within this stage of the research process (e.g. 'evidence of seeking out contradictory observations' because it was the "clearest operationalisation" of all the ways suggested for assessing the reliability and validity of data analysis" (Harden *et al.*, 1999b, p27). The seven chosen criteria are listed in table 6.2 on the next page.

During the review, reviewers were required to assess process evaluation reports against each of the criterion by answering specific questions. The style of the questions invited yes/no response options from reviewers. A modified and much shortened version of the Review Guidelines was used alongside the seven criteria to extract data for the description of evaluation methods and sample; the content of the intervention evaluated; and the types of processes assessed. Two members of the team (myself and Weston) extracted data and assessed quality independently and then met to establish a final agreed version.

Table 6.2: Criteria used to assess the quality of process evaluations

Criteria	Question	Response options	Additional guidance
1. An explicit theoretical framework and/or literature review	Did the report provide an explanation of, and justification for, the intervention and its evaluation using appropriate literature and/ or describe the theoretical framework used for the study?	Yes/No	This is intended to assess whether the research has demonstrated how it was informed by or linked to an existing body of knowledge.
2. Aims and objectives clearly stated	Did the report explicitly and clearly state the aims of the intervention and the evaluation?	Yes/No	None
3. A clear description of context	Did the report adequately describe the context of the intervention and the evaluation (e.g. intervention setting, target group)?	Yes/No	This is intended to assess whether all the factors that could be important in interpreting the results of the evaluation had been considered e.g. intervention setting, target group. Ideally there should also be some critical reflection on the evaluators' position and any possible consequences of this for the results.
4. A clear description of the sample and how it was recruited.	Did the report provide adequate details of the sample used to evaluate the intervention and how the sample was recruited?	Yes/No	This should include presentation of socio-demographic data and data on other salient factors such as descriptions of high risk groups.
5. A clear description of methods used to collect and analyse data.	Did the report provide an adequate description of the methods used in the study including its overall research framework, methods used to collect data and methods of data analysis?	Yes/No	None
6. Analysis of data by more than one researcher	Were the data analysed by more than one researcher?	Yes/No	None
7. Inclusion of sufficient original data to mediate between evidence and interpretation	Did the report present sufficient data in the form of, for example, data tables, direct quotations from interviews or focus groups, data from observation, to enable the reader to see that the results and conclusions are grounded in the data? Could a clear path be identified between the data and the interpretation and conclusions?	Yes/No	None

Additional guidance was not always provided to help reviewers assess whether aspects of process evaluation methods and results were 'clear', 'adequate' or 'sufficient'. This reflected the lack of, or inconsistent, detail provided in the original

tools. For example, within the 'data collection and data analysis methods' section in table 6.1, only Cobb and Hagemaster (1987) demand details on how entry was gained to participants and only Blaxter (1996) demands evidence of systematic data collection and record keeping. When additional guidance was provided it adapted the general principles of good practice embodied across the original tools for process evaluations. For example, in judging whether a study report 'provided an explicit theoretical framework/and or literature review', the four sets of criteria to assess the quality of 'qualitative' research presented in table 6.1 considered it important that a study was informed by an existing body of knowledge on the phenomenon under study and on the methods to be used to study it (see row 1 in table 6.1). Reviewers were therefore asked to consider whether the study report provided 'an explanation of, and justification for, the intervention and its evaluation using appropriate literature/ and or specified the theoretical framework used to guide the study' in order to demonstrate how the study 'was informed by or linked to an existing body of knowledge' (see row 1 in table 6.1).

In many systematic reviews it is usual practice to exclude or weight studies according to quality. As a result, the findings of poor quality studies are excluded or given less weight in the synthesis stage of a review. However, in this review the findings of all studies were used within the synthesis stage and subsequently informed the conclusions of the review about the appropriateness of peer-delivered health promotion. This was because we were not confident that the quality criteria used were the 'right' ones. The debates evident in the literature about assessing the quality of 'qualitative' research or process evaluations indicated that no-one else knew what the right criteria were either. In contrast to the situation for trials, there is no knowledge about what the 'fatal flaws' might be in process evaluations or 'qualitative' research (i.e. methodological problems which may cast serious doubt on trustworthiness of the study findings).

To our knowledge no other research group had, at this time, tried to assess the quality of process evaluations within a systematic review so our attempt was essentially a pilot exercise. In Harden *et al.* (1999b, p 26) we noted that "the process evaluations and their results were mapped against several quality criteria" with the aim "to provide the reader with a synthesis, within an explicit framework of methodological quality, of the findings of the process evaluations and their implications for developing and implementing peer-delivered health promotion for young people and the accessibility and appropriateness of peer-delivered health promotion". In practice this consisted of i) a commentary on the overall quality of the 16 process evaluations including, for example, the number of studies meeting each of the criteria assessed by the tool; ii) a synthesis of the findings of the process evaluations in the form of a structured narrative describing what the process evaluations collectively revealed (and did not reveal) about factors influencing the development, delivery and acceptability of interventions; and iii) detailed summaries of each individual process evaluation covering aims of the study and the intervention evaluated, methods used, findings and a short commentary on quality.

The draft report of the review was delivered one year after work on the review started in August 1999. We were excited about what we had achieved and I presented a methodological paper to the annual Cochrane Colloquium within a set of papers about 'qualitative' research and systematic reviews. This was the first time the Colloquium had accepted such a set of papers. There were five papers in the session including one by Oliver on the review of smoking cessation programmes described earlier (Oliver *et al.*, 1999b). It was at this point that we realised how far ahead we were in terms of developing methods; of the remaining three papers in the session, only one presented a worked example of a review that included 'qualitative' research (Roberts *et al.*, 1999). (The other two papers presented ideas about why it might be important to include 'qualitative' studies in systematic

reviews.) Peer and policy referee of the draft review report confirmed the innovative nature of the methods (e.g. "this review represents progress in terms of paying equal attention to outcome and process evaluations") and linked these innovations to the usefulness of the review product (e.g. "[the] inclusion of process evaluations, is welcome and adds a considerable amount of detail and depth to the findings of the review"). The full review was finally published in November 1999 and details on its methods were subsequently published as a journal article in 2001 (Harden *et al.*, 2001a). The review has since been used as an example of how to include process evaluations in systematic reviews in several major texts about methodological developments in research synthesis (e.g. Dixon-Woods *et al.*, 2004a; Kahn *et al.*, 2001; Mays *et al.*, 2001; Wallace *et al.*, 2004).

(iv) From process evaluations to studies of young people's 'views'

The next review in our 1998 to 2001 programme of work for the DH was on the barriers to, and facilitators of, health behaviour change amongst young people. Such a review would "draw on studies of health promotion interventions, but would also encompass a wider literature of studies, including 'qualitative' and needs assessment studies and survey data"[5]. I was a keen advocate for this review as I had previously researched barriers and facilitators in relation to young people's sexual health and was intrigued as to how one might go about reviewing this kind of research in a systematic way. Such a review also offered the opportunity to build on an earlier piece of EPPI-Centre work which had described different types of research on the "needs and views of young people with respect to their health and the range of interventions undertaken" (Peersman, 1996, p 1). The DH chose the barriers and facilitators review because it: a) was likely to support the 'Our Healthier Nation' strategy set out in the 1998 Green Paper (with its focus on social, economic

and social factors and their interaction with behaviour change); b) would have a focus on socially excluded young people (a focus on socially excluded groups was a strong theme in government policy and reducing inequalities in health a key goal); and c) would enable the scope of the databases at the EPPI-Centre to extend to a wider range of literature[6]. We were asked to draw up a full proposal for policy and external peer referee.

When we discussed the proposal amongst the team, Oakley argued that the underlying policy question for the review - 'why do so many young people engage in unhealthy behaviours given what is known about the short and long term consequences of such behaviours?' – should drive the review methods[7]. She suggested that we a) identify particular health topics to focus on; b) conduct literature searches in these topic areas; and c) sift out those studies which have findings relevant to the question 'why don't young people change their behaviour?'

A full proposal was drawn up to address two main questions: 1) 'What is known from research about the factors which promote or hinder young people's health behaviour change across a number of health topics/settings?'; and 2) 'How can the conclusions from this research improve the efficacy of health promotion interventions for young people?' The health topics suggested were physical activity, healthy eating and mental health in line with priority targets listed in the Green Paper. We proposed to include a range of primary research studies within two main categories: evaluations of health promotion interventions (from which information would be collected about why interventions failed, succeeded, or apparently had no impact), or other types of studies, including needs assessment, surveys, and

---

[5] Detailed in a proposal from the EPPI-Centre to the Department of Health.
[6] Letter from Sandra Williams of the DH Research and Development Division to Ann Oakley dated 7th February 1999.
[7] Note to Weston and Harden from Oakley 6th April 1999.

smaller-scale 'qualitative' studies (from which information would be collected about the factors which promote or hinder young people's health behaviours).

This proposal produced three separate reviews on three different topic areas: mental health; physical activity; and healthy eating. Work started on the mental health review in September 1999. By this time we had lost Ros Weston, which left myself, Oakley and Oliver to plan the initial design for the review and establish its scope and boundaries. The first tasks were to re-work the overall proposal outlined above specifically for mental health; set the scope of the mapping stage of the review by developing a set of inclusion and exclusion criteria; and run literature searches. A key challenge was to think about which types of studies would yield relevant findings for answering the review question 'What is know about the factors which promote or hinder good mental health amongst young people, especially those from socially excluded groups?' The potential scope of this review was huge. The term 'mental health' is ill-defined, but it is used nonetheless as an umbrella term to encompass all types of mental illness as well as concepts such as self-esteem and emotional skills. Moreover the proposal had suggested that a wide range of study types would be included in this review series. Understandably we were nervous about becoming overwhelmed by the task and we were keen to put some boundaries in place without compromising on research quality or policy needs.

The concept of 'positive mental health' or 'mental well-being' was employed in the review to group the outcomes of interest. This group would include the absence of mental illness, but go beyond to encompass 'resources' for reaching one's full potential (e.g. the ability to initiate, develop and sustain mutually satisfying personal relationships). The team hypothesised that factors which hinder good mental health ('barriers') or promote it ('facilitators') could be: interventions to promote good mental health shown to be effective or ineffective or harmful; interventions shown to be appropriate or inappropriate; or social, cultural, psychological or structural factors

associated with good mental health or mental health problems. It therefore followed that the review should be limited to particular study types: outcome evaluations to determine the effects of interventions; process evaluations to determine appropriateness; and 'non-intervention' or descriptive studies which can highlight the types of factors which are associated with good mental health or mental health problems. The latter category was anticipated to be quite a mixed bag including cross-sectional surveys or retrospective and prospective cohort studies. Having undertaken this conceptual and theoretical thinking it was time for us to stop working in the abstract and to engage with the literature.

Full searches were implemented and 11,638 citations were screened according to whether they reported a relevant study type on the relevant topic. By December 2000 it was becoming clear that the number of citations meeting the inclusion criteria far exceeded the team's expectations. More than 2000 studies had been labelled 'include' on the basis of their title and/or abstract. A meeting of the steering group for the programme of work was held in January 2000 and we asked for advice on how to manage the review scope. We presented two main options to the steering group: a focus on one of the many mental health outcomes or restricting inclusion to studies with particular methodological attributes. It was agreed by the steering group that we should include outcome evaluations from any country but only include 'non-intervention' research carried out in the UK. This was considered a pragmatic way of cutting down the work that could sensibly be defended in the review product. We argued that the strength of non-intervention studies lay in their ability to describe barriers and facilitators within specific countries that could be contrasted with the barriers and facilitators studied in international intervention research.

The 11,638 citations were re-screened accordingly and a descriptive map was eventually carried out on 345 studies. All studies were coded according to a standardised strategy and the results were presented to another meeting of the

programme of work steering group in May 2000, and in a report to the DH in June 2000. This time we needed direction for prioritising a sub-set of studies for in-depth review as their searches had uncovered 187 intervention studies (and a further 25 reviews of intervention studies) and 133 non-intervention studies. At the steering group meeting, members recommended that we should only review in-depth intervention studies which had not already been included in systematic reviews, but suggested that we consult the DH with respect to priority mental health outcomes. For the non-intervention studies, the steering group suggested that we focus on 'qualitative' studies of people's perspectives and experiences, given the experience of SSRU in conducting this type of research.

We explored the content of the category of research that we were calling 'non-intervention' studies in more detail. Many of these studies measured the statistical association between mental health outcomes and possible causal factors (e.g. the relationship between socio-economic status and depression). A small number, however, had conducted surveys of young people to gather their perspectives and experiences of mental health. These studies tended to focus on mental health not further specified rather than particular types of mental health outcomes such as depression, suicide or self-esteem. Like the steering group, we were particularly drawn to this set of studies because we knew that intervention research rarely involved the views of intended recipients in intervention development and evaluation (Harden and Oliver, 2001). A focus on these studies would also mean that we could build on our work to include 'qualitative' and other types of process evaluations in the review of peer-delivered health promotion. We consulted with the DH to check that focusing on studies that examined young people's perspectives and experiences would be useful from a policy point of view. The DH was happy for us to proceed on this basis as it reflected their recently pledged commitment to involve

the public in the development and delivery of services (Department of Health, 1999b).

In July and August 2000 we developed the procedures for, and then undertook an in-depth review of studies of young people's views about mental health (alongside relevant intervention research). This was a frantic time but three new members had by this time arrived. Rebecca Rees, Jonathan Shepherd and Ginny Brunton all brought additional disciplines, skills and experience to the team. Shepherd and Brunton had both worked collaboratively with the EPPI-Centre for a number of years: Shepherd had led a Cochrane review on the prevention of cervical cancer with Ros Weston and Greet Peersman, and Brunton was part of the team at McMaster University in Canada which jointly coordinated the Cochrane Health Promotion and Public Health Field with the EPPI-Centre. Brunton had previously trained as a nurse and joined the EPPI-Centre after a number of years conducting systematic reviews to inform public health policy and practice for the Hamilton province in Canada. Shepherd had degrees in geography and health promotion and had also conducted a trial of a peer-led intervention to promote HIV-related sexual health amongst gay men. Rees joined SSRU initially to work on a project to develop an evidence-based approach to involving consumers in research and development agenda-setting in the NHS. Like Oliver and Peersman, Rees had started her career in the biological sciences, but had gone on to undertake a masters degree in social research methods and statistics. Prior to joining SSRU, she had worked for a charitable organisation which co-ordinated the Complementary Medicine Field of the Cochrane Collaboration. She had also spent some time campaigning and researching for the Women's Environmental Health Network.

The experience and enthusiasm of the team were important for the task ahead. With Oakley and Oliver helping to thrash out any problems as they arose, Brunton and Shepherd worked on the in-depth review of intervention research, whilst Rees and I focused our efforts on the in-depth review of studies of young people's perspectives and experiences (which we referred to as 'views' studies for short). Initially we spent some time working out what a 'views' study was and how it might be recognised. In the map we had applied the 'views' code but had not provided detailed instructions for using the code. (Instructions were to apply it to "studies which directly ask young people for their own views on mental health".) We were therefore unsure about whether they had applied the code in the same way. We re-examined all studies that had been coded as 'views' or 'qualitative' or 'survey'. A particular set of studies that caused problems with coding for the map were those that had asked young people for their views, but translated these responses into attitudinal or other types of variables to enter into statistical models to trace causal pathways to mental health outcomes. Sometimes these studies had been coded 'views', sometimes they had not. One strategy we considered to avoid this problem was to restrict the definition (and inclusion) of 'views' studies to those which used 'qualitative' methods of data collection.

We were reluctant to do this, however, because it might reinforce traditional and unhelpful distinctions between 'quantitative' and 'qualitative' research. Instead we proposed that a study, to be considered as one of young people's views about mental health, had to: (i) examine young people's attitudes, opinions, beliefs, feelings, understanding or experiences, rather than their health status, behaviour or factual knowledge about mental health issues; (ii) access views about: young people's definitions of and/or ideas about mental health, their ideas about factors influencing their own or other young people's mental health and about ways of promoting this; (iii) privilege young people's views: studies had to present young

people's views directly as data that are valuable and interesting in themselves, rather than as a route to generating variables to be tested in a predictive or causal model (e.g. measuring a range of attitudes or experiences to see whether/how these predict mental health status).

Now we were clearer about what a 'views' study was, Rees and I set about designing a data extraction and quality assessment framework. I had already drafted a framework based on the 'Review Guidelines' for outcome and process evaluations (Peersman *et al.*, 1997) and the quality criteria used to assess process evaluations in the review of peer-delivered health promotion. We both applied this initial draft to one views study independently and then met to compare problems encountered and to work out modifications. The final version, entitled 'Data extraction and quality assessment guidelines for research examining young people's views' required reviewers to record descriptive information in a standardised way within six sections before they were asked to judge the quality of the study in a final seventh section. The six sections collecting descriptive information contained 56 questions which covered study identification (10 items) (e.g. bibliographic details); support for the study (two items) (e.g. source of funding); the aims and context of the study[8] (10 items) (e.g. where the study was carried out; whether the study was informed by previous research); sampling and sample (16 items) (e.g. sampling and recruitment; characteristics of sample); data collection and analysis (12 items) (e.g. methods used; reliability and validity); and study findings (six items) (e.g. the barriers or facilitators identified by young people).

In the quality assessment section of the guidelines (shown in table 6.3) all but one of the criteria that we previously employed in the review of peer-delivered health

---

[8] Context was defined in the guidelines as the "specific circumstances under which the research was developed, carried out and completed. Such circumstances may impact on the findings of the study and can provide information on the applicability and relevance of the findings to other situations".

Table 6.3: Criteria and associated guidance for assessing the quality of studies of young people's perspectives and experiences

Criteria	Question	Response options	Additional guidance
1. An explicit theoretical framework and/or literature review	Does the study give an explicit account of a theoretical framework and/or include a literature review?	Yes/No	*Consider your answer to:* Is the study informed by previous research and/or a specific theoretical framework?
2. Aims and objectives clearly stated	Did the report explicitly and clearly state the aims of the study?	**Yes/No**	*Consider your answer to:* What were the aims of the study?
3. A clear description of context	Did the report adequately describe the context of the study?	Yes/No	*Consider your answer to:* Country in which study was carried out; What were the aims of the study? Who was involved in identifying the aims of the research?; What was the rationale for undertaking the study? (what reasons do the authors give for conducting this piece of research?, why was it considered important to undertake this study?, why was this particular topic/group of people/setting the focus of the investigation); Study topic area; Did the research focus on a particular group of young people?; Who carried out the research?; Did those involved in carrying out the research discuss or reflect upon their potential impact upon the findings of the study?
4. A clear description of the sample and how it was recruited.	Did the report provide clear details of the sample and how the sample was recruited?	***Yes/No***	*Consider you answer to:* Characteristics of the study sample; Do authors report reasons why not all those selected for the sample provided data?; and (if applicable) Do authors report the number of people who dropped out of the study?
5. A clear description of methods used to collect and analyse data.	Did the report provide a clear description of the methods used in the study including methods used to collect data and methods of data analysis?	**Yes/No**	*Consider you answer to:* Do the authors provide enough details on methods of data collection and analysis to be able to replicate the study?
6. Attempts made to establish the reliability or validity of data analysis	Are there attempts made to establish the reliability and/or validity of the data analysis?	Yes/No	***Consider you answer to:*** Do the authors describe any methods for ensuring the reliability and validity of the data analysis? (e.g. using more the one researcher to analyse data, feeding back analysis of data to participants, looking for negative cases)
7. Inclusion of sufficient original data to mediate between evidence and interpretation	Were sufficient original data included to mediate between data and interpretation?	**Yes/No**	***Consider you answer to:*** Are sufficient data presented to illustrate the themes presented by the author?

promotion were included. One of the peer referees of the review of peer-delivered health promotion had argued against the criterion 'analysis of data by more than one researcher' because it is unknown whether using more than one researcher to carry out data analysis is likely to lead to more trustworthy findings. Indeed this referee suggested that using more than one researcher might introduce 'error' if the second researcher had not been out in the field actually collecting the data. With this in mind, and because this criterion is only one of a number of advocated strategies for establishing the reliability and/or validity of data analysis, it was replaced with the more generic 'attempts made to establish the reliability or validity of data analysis'.

Although the criteria were similar, the procedures for applying the criteria were a little different. As noted above we only developed very basic guidance for assessing the quality of the process evaluations. We therefore provided more extensive and systematic guidance to help reviewers in their assessments of the quality of young people's views studies, prompting them to refer back to the relevant descriptive information they had recorded in earlier sections of the guidelines. The guidance was especially important given the diverse backgrounds of team members. The main way in which the guidance was developed was through lengthy discussion and debate amongst the team after trying out the criteria.

All of the above development work was completed by the middle of July 2000. The next six weeks were spent reviewing 12 views studies in depth, synthesising their results and conducting a 'cross-study' synthesis to bring together the views studies with the intervention synthesis. (Synthesis methods are described elsewhere, see Harden *et al.*, 2004; Oliver *et al.*, 2005.) A first step was to prepare structured summaries of each individual study covering: aims and rationale for the study, sample recruitment and characteristics, data collection and analysis methods,

findings, and a short commentary on study quality. In addition, two types of 'evidence table' were prepared: one detailing the methods used in each study and reviewers' judgments about the methodological quality of the study, and one detailing authors' and reviewers' conclusions about the findings of studies, alongside study aims and sample characteristics, and reviewers' judgments about quality and study findings.

Rees and I laid out the draft structured summaries and evidence tables on a big table and puzzled over how to synthesise the findings from the 12 views studies. The studies varied in methodological quality: only two studies had m*et al*l seven quality criteria; two had met six; one had met five; three had met four; one had met three; one had met two; and two studies had only met one of the seven quality criteria. We did not feel that there were good reasons to exclude any of the studies from the synthesis for the same reasons outlined earlier: we were not confident that the quality criteria used were the 'right' ones and no-one else knew what the right criteria were either. We were also reluctant to lose any of the studies as there were so few to start with. Something else was bothering us about the studies, though. Surveying young people's views about mental health was common to all the studies (and most of the study authors offered an explicit rationale for the importance of attending to young people's views), but we noted many differences between the studies which we felt that it was important to take account of in our analysis and synthesis. Three things troubled us in particular.

Firstly, studies differed in their substantive starting point. Some studies focused on one particular aspect of mental health such as self-esteem, some focused on mental health as a whole, whilst others asked young people what they were worried or concerned about. Another dimension here was that some studies had been undertaken especially to inform the development of an intervention whilst other

studies appeared to have been undertaken to contribute to knowledge in this area. Secondly, not all studies had addressed directly the team's review question, 'what do young people see as the barriers to, and facilitators of, good mental health?' Some studies did not go any further than asking young people about their attitudes to mental health or about what the term mental health meant to them. In other studies, authors had inferred barriers and facilitators from what young people said rather than ask them directly. Thirdly, despite study authors' emphasis on listening to young people, there was little evidence in the studies that young people had been actively involved in the design or implementation of studies. For example, study authors reported that consent had been sought in only three studies and only half of the studies reported making data anonymous or assuring young people that responses would be treated in confidence. To move forward we did two things: 1) considered each study in terms of its contribution to the review question 'what do young people see as the barriers to, and facilitators of, good mental health?'; and 2) flagged up the methodological problems in the studies not captured by the quality assessment criteria in the review report.

After the draft report of the young people and mental health review was completed in September 2000, work immediately started on the physical activity and healthy eating reviews. All three reviews used the same approach and methods and were eventually published in September/October 2001. The reports were generally well received by policy-makers and topic-expert academics who viewed the inclusion of 'qualitative' research to have added an important new dimension to the literatures on mental health, physical activity and healthy eating. Another big impact of the reports was within the world of systematic reviews. The reviews were some of the first to integrate 'qualitative' research alongside trials and were considered internationally to have represented a significant breakthrough in systematic review methodology. One disappointing finding from all this review work was the poor

quality of much of the 'qualitative' research in health promotion and public health. Many of the 35 studies we had reviewed failed to meet basic standards of reporting quality. For example, less than one fifth of studies had employed strategies within data analysis to increase the validity of study findings. Moreover, as indicated above, the studies also suffered a number of other methodological problems not captured by our tool. The quality of the research was therefore one factor that was crucial for the next stage of development.

(v) Systematic reviews in education and further developments in HP&PH

Whilst the review series on young people was underway, we had also (with other colleagues) begun a new programme of work to develop systematic reviews in education. With this programme we planned to translate, adapt and extend the methods and tools we had developed for doing systematic reviews in health promotion for education reviews. In contrast to the health promotion programme, this time we were developing methods for other groups to use in their reviews rather than for our own reviews. Although we were used to working in a multi-disciplinary team, the education review groups did not necessarily share the same principles or understandings about research as us. This was particularly evident around study type. Our way of classifying research into categories such as intervention and non-intervention research, outcome and process evaluations, and studies of people's views was not thought to be valuable by the education review groups. For example, they did not see the various strategies and programmes within education as 'interventions', viewed randomised controlled trials as largely inappropriate, and wanted to apply categories such as 'case study' or 'qualitative' research.

Taking some of these views on board, we offered the education review groups a broad-based review system which would apply to *any* type of study regardless of

how it was classified. We developed a set of generic data extraction and quality assessment guidelines using the set we had developed previously to apply to studies of young people's views in the health promotion reviews. (The items included within the guidelines developed for studies of young people's views were actually quite generic already and needed relatively little modification.) Review groups were advised to develop their own review-specific guidelines if they needed to code additional aspects of studies not covered by the generic guidelines.

As the first drafts of the education reviews came into the EPPI-Centre in early 2002 we spotted an unintended consequence of the move towards generic data extraction and quality assessment guidelines. We had inadvertently broken the link between research/review question, data extraction and quality assessment. Within the reviews in health promotion, study types were matched closely to the review question. We usually specified in advance the kinds of studies which would be the most appropriate to answer particular review questions (e.g. trials to answer questions about effectiveness) and, for outcome evaluations, quality assessment was linked to the review questions. In the education review groups, however, a desire to be as inclusive as possible meant that nearly all of the review groups had not considered the issue of the appropriateness of the study design to the research question in their inclusion criteria or quality assessment procedures. A big problem appeared to be confusion between study design and methods of data collection and analysis. Review groups wanted to include studies that had used 'qualitative' methods of data collection and analysis but had not considered whether or not these studies had used appropriate designs for addressing their review questions. Another issue to arise from such an over-inclusive approach was that studies were not always a close match to the phenomenon under review (e.g. context, population).

As a solution to this, a framework called 'weight of evidence' was devised. This was led by David Gough, a psychologist with a research and teaching career in the social welfare field. He had joined SSRU in 1998 and was a leading figure in the establishment of the EPPI-Centre as the centre for evidence-informed policy and practice in education in 2000. Influenced by the work of Slavin (1995) on best evidence synthesis, Gough proposed a three dimensional framework for judging the weight of evidence a study could contribute to a particular review. (See Gough, forthcoming for a detailed overview.) The first dimension – the quality of execution of the study in its own terms – requires reviewers to judge the quality of a study according to the usual standards associated with that particular type of study. For example if it is a randomised controlled trial a reviewer would judge whether it has been properly designed, executed and analysed. The second dimension requires reviewers to judge the appropriateness of the study design and analysis for answering the review question under study. The third dimension asks reviewers to judge how well matched the study is to the focus of the review in terms of the topic under study and its operationalisation, the sample and population, the context, and/or any measures used. Each study in a review is judged as 'high', 'medium' or 'low' on each of the three dimensions. A final step involves reviewers making an overall judgement about whether each study contributes a high, medium or low weight of evidence to the review.

By the time the first education reviews had been completed, three more health promotion reviews were underway for the DH: one on children and physical activity; another on children and healthy eating; and one on HIV health promotion for men who have sex with men (MSM). For each review the DH wanted us to use the same approach as in the young people review series because they wanted a range of different questions answered. For example the review of HIV health promotion for MSM was commissioned at a time when this group were still at greatest risk of

acquiring HIV infection in the UK. Recent surveys had shown that self-reported risky sexual behaviour was on the increase amongst gay men, reversing the trend towards safe sex revealed by surveys in the early years of the HIV epidemic (Dodds *et al.*, 2001; Johnson *et al.*, 2001). Policy-makers at the DH wanted to know why this reversal was happening (e.g. have advances in treatment for HIV changed attitudes and led to complacency?) and what could be done to stop it (e.g. what is the effectiveness of community level interventions?)[9]. Illuminating recent trends in risk behaviour was therefore a key policy question which 'qualitative' and other types of studies of the views of men who have sex with men could help to answer.

We fed back into these three new health promotion reviews our experiences of working with education review groups. Rather than using the weight of evidence framework in its entirety though, we worked in particular with the second dimension and tried to link our quality assessment criteria to the relevant questions of the reviews. As a consequence our earlier set of seven criteria grew to a total of 12 (table 6.4). One of the major changes was the addition of a section on the appropriateness of methods for studying people's perspectives and experiences. We hoped that these criteria would pick up on the quality issues not covered by the previous version of the tool which had focused largely on the quality of reporting. We had noted three issues in particular i) a lack of pilot work prior to finalising data collection tools to ensure that questions and response categories were meaningful to participants; ii) the use of pre-defined coding strategies and/or a lack of detail on methods of data analysis which meant that it was difficult to tell whether findings were grounded in people's perspectives and experiences; and iii) a lack of attention to the active involvement of participants in the research. A final step in the quality assessment process using the tool described in table 6.4 was to assign each study

---

[9] Recorded in notes from a meeting between the Department of Health and the EPPI-Centre on 3rd May 2002.

a weight of evidence. Reviewers were asked 'What weight of evidence would you give this study in terms of whether its findings are really rooted in the perspectives of children/MSM?' (Hints for reviewers were, in the physical activity review, whether they believed the study to be rooted in what participants think or the researchers think, or in the HIV-health promotion review, whether the study could have distorted, misrepresented or failed to pick up people's views.)

Table 6.4: Criteria and guidance for assessing the quality of studies in reviews about the views of children and those of men who have sex with men (MSM).

Criteria	Question	Response options	Additional guidance
Quality of reporting			
1) Aims and objectives clearly reported	Are the aims of the study clearly reported?	Yes/No	*Consider your answer to questions*: What are the broad aims of the study?; and What are the study research questions and/or hypotheses?
2) Adequate description of the context of the research	Is the context of the study adequately described?	Yes/No	*Consider your answer to questions*: Why was this study done at this point in time, in those contexts and with those people or institutions?; Was the study informed by, or linked to an existing body of empirical and/or theoretical research?; Which of the following groups were consulted in working out the aims to be addressed in the study?; Do the authors report how the study was funded?; When was the study carried out?
3) Adequate description of the sample and how it was identified and recruited	Is there an adequate description of the sample used in the study and how the sample was identified and recruited?	Yes/No	*Consider your answer to questions*: Are the authors trying to produce findings that are representative of a given population?; What is the sampling frame (if any) from which the participants are chosen?; Which methods does the study use to select people, or groups of people (from the sampling frame)?; Planned sample size; Which methods are used to recruit people into the study?; Were any incentives provided to recruit people into the study?; Was consent sought?; No of participants? Age, sex, socio-economic status, ethnicity and other characteristics of sample?
4) Adequate description of data collection methods	Is there an adequate description of the methods used in the study to collect data?	Yes/No	*Consider your answer to questions*: Which methods were used to collect the data; Details of data collection methods or tools; Who collected the data?; Do the authors describe the setting where the data were collected?; Are there other important features of the data collection procedures?
5) Adequate description of data analysis methods	Is there an adequate description of the methods of data analysis?	Yes/No	*Consider your answer to questions*: Which methods were used to analyse the data?; What statistical methods if any, were used in the analysis?; Who carried out the data analysis?

Table 6.4 (cont'd): Criteria and guidance for assessing the quality of studies in reviews about the views of children and those of men who have sex with men (MSM).

Criteria	Question	Response options	Additional guidance
Strategies for establishing reliability and validity			
6) Sufficient attempts to establish the reliability of data collection tools.	Have sufficient attempts been made to establish the reliability of data collection methods and tools?	a) Yes, good b) Yes, some attempt c) Yes, minimal attempt d) No, none	Do the authors describe any ways they have addressed the reliability of their data collection tools/methods? (e.g. test - re-test methods.)
7) Sufficient attempts to establish the validity of data collection tools.	Have sufficient attempts been made to establish the validity of data collection tools and methods?	a) Yes, good b) Yes, some attempt c) Yes, minimal attempt d) No, none	Do the authors describe any ways they have addressed the validity of their data collection tools/methods? (e.g. mention previous validation of tools, published version of tools, involvement of target population
8) Sufficient attempts to establish the reliability of the data analysis methods.	Have sufficient attempts been made to establish the reliability of data analysis?	a) Yes, good b) Yes, some attempt c) Yes, minimal attempt d) No, none	Do the authors describe any ways they have addressed the reliability of data analysis? (e.g. using more than one researcher to analyse data, use of a software package.)
9) Sufficient attempts to establish the validity of data analysis methods.	Have sufficient attempts been made to establish the validity of data analysis?	a) Yes, good b) Yes, some attempt c) Yes, minimal attempt d) No, none	Do the authors describe any ways they have addressed the validity of data analysis? (e.g. searching for negative cases; checking results with participants.)
Appropriateness of methods for studying children's/MSM's perspectives and experiences			
10) Appropriate data collection methods used for helping people to express their views.	Does this study use appropriate data collection methods for helping children/MSM to express their views?	a) Yes b) Partially c) No d) Can't tell	None given
11) Appropriate methods used for ensuring data analysis grounded in people's views.	Does this study use appropriate data analysis methods to help ensure that study findings are grounded in the perspectives of children/MSM?	a) Yes b) Partially c) No d) Can't tell	None given
12) Appropriate active involvement of study population in design or conduct of the study.	Were children/MSM actively involved in the design/conduct of the study?	a) Yes b) Partially c) No d) Can't tell	None given

Despite using this more detailed quality assessment tool, we did not exclude any studies on the basis of quality in the children and physical activity review or the children and healthy eating review. It was not until the HIV health promotion review that we made a decision to exclude studies which had been judged to be a low

weight of evidence. This decision was prompted by the nature of the claims that had been made in one of the low quality studies in particular: a study of deaf gay men, safer sex and HIV. Reeves (1999) aimed to identify the needs of deaf gay men in relation to safer sex and HIV and concluded that the study findings indicated "an urgent need for more effective HIV health promotion targeted specifically at deaf gay men" (p27). Whilst this may well be the case, we were very concerned about the findings leading to this conclusion. Some of these findings were based on data from in-depth interviews. For example the author argues that safer sex campaigns have not reached deaf gay men because they are largely written in English and, as a result, there are high levels of unsafe sex and "a dangerously low level of knowledge of safer sex" (p17). However, because the author did not report on how the interviews were conducted and analysed, reviewers were reluctant to take these findings at face value. It was not clear whether the findings emerged from a rigorous analysis of the interview data or whether the author had been highly selective by using quotes to illustrate preconceived ideas or understandings. Another issue that compromised the findings of the interviews was a discrepancy surrounding the number of people who were interviewed. The author reports that six people took part in the interviews, but quotes are presented from a total of 17 different people. It was not clear why there was such a discrepancy.

We were also concerned about other findings that led to the conclusion that deaf gay men needed urgent intervention. These were derived from comparing the views of deaf gay men collected in this study via anonymous self-completion questionnaires with those of hearing gay men collected in another study using questionnaires administered face to face by a researcher. The study author highlighted many differences between these two groups in, for example, their thoughts during or before unsafe sexual encounters (e.g. 14% of deaf men compared with 4% of hearing men said that their thoughts included 'He'll think I've

got HIV if I ask to use a condom'). However, the differences were in fact very small, no assessment of the significance of the differences was undertaken, and the methods used to collect the data were different in the two groups. It was therefore possible that these differences were not real but due to chance or due to differences in methodology.

The three other excluded studies judged to be of low weight of evidence had similar problems. Docherty (2002) explored influences on the sexual practices of HIV positive men and concluded that "quality of life factors", such as taking risks for a more fulfilled sex life, impact upon sexual practices. Patel *et al.* (1999) explored the sexual health needs of South Asian men who have sex with men and drew several implications for health promotion practice (e.g. respond to sexual behaviour not just sexual identity, understand different cultural expressions of sexuality). Ward (2002) examined whether health promotion service provision was inclusive of the needs and rights of HIV positive men and identified likes and dislikes about services, how they could be improved, views on peer support, and how well services had involved them in service design and delivery. Reviewers' comments about each of these studies all emphasised the non-existent description of analysis methods, failure to employ (or mention) techniques to enhance the rigour of the analysis, and a lack of evidence from participants accounts to support the interpretations of the authors.

6.4 Discussion

The analysis reported in this chapter revealed several inter-related factors that encouraged the development of a new and unique tool to assess the quality of 'qualitative' and other types of research. The intersection between policy needs and our interests as a research team was clearly one of the most significant drivers of the development. The team wanted to work on the methodological challenges

presented by including 'qualitative' research in systematic reviews and policy needs offered continued funding to do reviews that required the inclusion of 'qualitative' research. This enabled us to take forward, and build on, progress made on one review to the next. This intersection also provided a real, rather than an abstract, context in which to pursue methodological issues because our task was to assess quality to provide reliable answers to policy questions.

In contrast to fears that greater involvement of policy-makers in research will be a vehicle for political control of the research agenda (Hammersley, 2001; Vulliamy and Webb, 2001), the analysis reported in this chapter suggests that the relationship between a research team and policy-makers can be a collaborative and mutually beneficial one, with 'give' and 'take' on both sides. In our case, by including 'qualitative' and other types of research, we had tried to make our reviews as relevant as possible to policy-makers without compromising on our scientific principles. The benefit to the funder was the production of relevant research with the added bonus of methodological development to help facilitate more relevant research in the future. The benefit to the team was intellectual in nature. Systematic reviews have been characterised by some as 'boring', 'mechanistic', and 'derivative' research (Hammersley, 2001; MacLure, 2005; Schwandt, 1998). Whilst this is an unfair characterisation of the systematic review field as a whole, like all research, following procedures and methods in a rigorous and systematic way can be tedious (and no researcher would want their work to be described as 'boring' or 'mechanistic'). This is especially true perhaps if one is not an expert in the topic area of the review. Although the team at the EPPI-Centre had expertise in the general area of health promotion and public health, none of us were topic experts in peer education, mental health, physical activity or healthy eating. It is unlikely therefore that doing straightforward systematic reviews in these topic areas would have provided sufficient intellectual stimulation for the team. Undertaking

methodological development introduced intellectual motivation for the team, which in turn explains why the methodological development took place despite a lack of funding.

The different disciplines represented within the team were important for bringing to the methodological development a diversity of perspective. The non-hierarchical way in which we worked meant that this diversity of perspective was able to make a genuine contribution. However, it was not the multi-disciplinary nature of the team *per se* which was crucial. It was the multi-disciplinary nature combined with the shared principles and interests of the team for producing research that is both scientific and relevant to policy. This undoubtedly facilitated our progress because we were all focused on solving the policy problems under study rather than getting bogged down in discipline specific debates. Moreover, our team commitment to both RCTs *and* 'qualitative' research - an unusual position for social and other types of scientists in the UK and elsewhere - meant that we were also relatively free of the 'one-upmanship' usually associated with adopting either the 'quantitative' or 'qualitative' side of the paradigm divide. The location of the team within a social science research unit with an emphasis on policy-relevant research was important here in terms of providing both intellectual and social support for our position. Producing policy relevant research can cut across the divide between 'qualitative' and 'quantitative' approaches as it focuses attention on the research question as the driver for research method. In contrast to all previous tools to assess the quality of 'qualitative' research, methodology was a secondary issue in our tool. Our tool reflected a question-led perspective on research because we aimed to use it to assess the quality of 'qualitative' research in relation to the review question under study.

A final factor identified as influential in the development of the new tool was the quality of the research itself. We were surprised by the failure of many of the 'qualitative' and other types of studies of intervention processes and people's perspectives and experiences to meet the basic methodological standards we had assessed them against. However, this gave us plenty of material to work with to identify the range of problems that might undermine a study's ability to provide trustworthy answers to our policy questions. We were then able to build in an assessment of studies according to these problems in the later version of our tool. The quality of the research we reviewed was therefore crucial in helping us move the tool beyond a consideration of basic methodological standards to the assessment of quality in relation to a specific review question.

As noted in the introduction to this chapter, to date there has been little critical reflection on the processes involved in the development of tools to assess the quality of 'qualitative' research. A recent exception is a paper by Attree and Milton (2006) which reflects on the development of a tool they used in a systematic review of 'qualitative' research on young people's experiences of growing up in disadvantage. Like the analysis reported here, Attree and Milton (2006) found that their own knowledge, experience, and interests influenced the development and application of their tool. They report, for example, that their choice of quality criteria reflected their "concerns as experienced qualitative health researchers, and our ideas about the important elements of a 'sound' qualitative study" (p23). Such reflection can help a reader to assess whether they would find the tool useful given their particular purpose and perspective. Tool authors who do not undertake explicit reflection on how they were developed cannot offer such information to help readers and can give the impression that there is only one way to assess the quality of 'qualitative' research. Together with Attree and Milton (2006), the analysis reported

here adds a new dimension to the literature on assessing the quality of 'qualitative' research.

The findings of this analysis also contribute to the literature on how science operates and develops. Like the large body of research that has studied the social nature of science (e.g. Harding, 1991; Rose, 1994; Woolgar, 1991) this analysis has shown how the practice of social science is influenced by the interest of scientists and their wider social context. Others have also highlighted the use of multi-disciplinary teams as an important factor for development in substantive issues as well as methods (e.g. Fuqua *et al.,* 2004; Hulme and Toye, 2006; Massey *et al.*, 2006; Tishelman *et al.*, 1999). 'Applied' research - designed to meet practical requirements as opposed to research designed to contribute to a body of knowledge with no immediate practical application - has traditionally been afforded lower status than 'basic' research (Hammersley, 2000a). This analysis suggests that conducting research to address practical questions can offer a fresh perspective for methodological development. Although I would not claim that the methodological development analysed here represents anything so grand as a paradigm shift, the following description by Kuhn (1970, p91) of the "transition from normal to extraordinary research" captures nicely some of the features of the story of methodological development told in this chapter:

> "Confronted with anomaly or with crisis, scientists take a different attitude toward existing paradigms, and the nature of their research changes accordingly. The proliferation of competing articulations, the willingness to try anything, the expression of discontent, the recourse to philosophy and to debate over fundamentals, all these are symptoms of a transition from normal to extraordinary research" Kuhn (1970, p 91)

In our case the "anomaly" was to have the research question as a starting point rather than a particular methodology or tradition within the social sciences. Such a perspective jarred with much methodological writing in the social sciences, and was also one which appeared to be quite alien and even distasteful to the education review groups we worked with. They rejected a question-driven approach and wanted to retain the distinction between 'qualitative' and 'quantitative' research within their review tools. We, however, persisted in taking a critical stance towards both the 'quantitative' and the 'qualitative' paradigms, preferring to go back to first principles to debate, for example, whether it can ever be meaningful to categorise a particular study as either 'qualitative' or 'quantitative'. (With respect to 'paradigm shifts', it is interesting to note that question-led research has been identified as a feature of a new 'mixed methods' research paradigm. 'Mixed methods' research is defined as "the class of research where the researcher mixes or combines quantitative and qualitative research techniques, methods, approaches, concepts or language into a single study" (Johnson and Onwuegbuzie, 2004, p 17). Johnson and Onwuegbuzie (2004, p 18) suggest the 'mixed methods' paradigm - which is linked to a pragmatic system of philosophy that emphasises that "research methods should follow research questions in a way that offers the best chance to obtain useful answers" - as an alternative to the 'quantitative' or 'qualitative' paradigm.)

The findings of this analysis suggest that methodological innovation can be fostered by a multi-disciplinary team working with shared principles and interests, within continuous programmes of research seeking to answer applied or policy and practice questions. These findings, however, need to be considered in the context of the strengths and limitations of the method I used to analyse the development of the new tool. Using other key members of the team to check my reconstruction of events and to discuss the factors I had identified as influential was crucial for enhancing the rigour of the analysis. This process ironed out factual inaccuracies

and enriched my explanation for how the methodological development was achieved by the team. However, a limitation of the analysis was that it was solely informed by an 'insider' perspective. The resulting account of the methodological development may only be one interpretation of the events. An insider perspective may have led me to miss or discount particular events that others would see as highly relevant to the methodological development. For example, it is possible that I was overly influenced by a desire to cast the research team I was part of in a positive light and, as a consequence, may have missed relevant and illuminating points of tension within the team.

Future work of this type which uses an 'insider' perspective to analyse the work of research teams should consider involving a peer from outside of the team. Such a peer could review the analysis as it progresses. In situations like the one here, in which research teams share the same principles and interests, it might be valuable to choose a peer who has a different set of principles and interests. This could help to highlight factors which the research team may take for granted, discount, or overlook because they are outside shared understandings. Such a strategy may also be useful for the methodological development itself. In the case described in this chapter a different perspective may have been obtained from someone who identified as a 'qualitative' researcher. As already noted, we actively resisted identifying ourselves as either 'qualitative' or 'quantitative' researchers and took a critical view of paradigm differences. However, by concentrating on common elements of all scientific research – using methods to reduce bias and error - we may have inadvertently downplayed meaningful differences between different types of research.

One possibility is that we downplayed the importance of considering the findings as well as the methods when judging the quality of 'qualitative' research. The survey of

previous tools to assess the quality of 'qualitative' research reported in chapter five found that items requiring readers to examine the findings of 'qualitative' research accounted for a significant proportion of the items overall. Moreover, systematic reviewers found those tools that included items about findings to be more useful than tools dominated by items about methods. Items about findings included, for example, whether or not sufficient evidence had been provided to show that findings were grounded in the data and whether concepts or theory were well developed. In contrast to many previous tools, our new tool emphasised a reviewer focus on methods rather than findings. Incorporating items about findings could therefore be a valuable step for the next stage of development of our tool. The importance of findings also came to light in the third and final methodological study in this thesis. This third study - an analysis to examine the relationship between study quality and the findings of systematic reviews that include 'qualitative' research - is reported in the next chapter.

CHAPTER 7

Study 3: Does the quality of 'qualitative' studies affect the findings of systematic reviews?

7.1 Introduction

The issue of whether '*how* we know what we know' influences '*what* we know' is a very important one for both producers and users of research and is the underlying focus of this thesis. The question posed in the title of this chapter is, however, very difficult to answer. Methodologists have studied the question of whether the quality of a trial matters for several decades. A number of systematic reviews have found that poorly designed and executed trials tend to give different answers about the effects of interventions compared to good quality trials (e.g. Abraham *et al.*, 2004; Guyatt *et al.*, 2000; Moher *et al.*, 1998; Peersman *et al.*, 1999; Schulz *et al.*, 1995). Although it is not yet possible to predict the direction that these differences will take - poorly designed and/or executed trials may overestimate, underestimate or mask the effects of interventions (Britton *et al.*, 1998; Kunz and Oxman, 1998; MacLehose *et al.*, 2000) - all of this work has increased our understanding about the different sorts of bias and error that can be introduced in a trial, and there is growing consensus about what the 'fatal flaws' might be (e.g. Juni *et al.*, 2001a,b). Moreover, this work forms a significant body of empirical evidence to support the exclusion of poor quality trials from reviews.

Trials are a specific type of study that answers a specific research question about the effects of an intervention. For other study types and research questions, whether or not to exclude poor quality studies is a matter of much debate, especially

within the literature on systematic reviews that include 'qualitative' research (Attree and Milton, 2006; Barbour and Barbour, 2003; Dixon-Woods *et al.*, 2004b; Noblit and Hare, 1988; Sandelowski *et al.*, 1997). There is, however, very little work examining the relationship of study quality to review findings which could inform this debate. Although there is a body of research emerging for studies addressing diagnostic questions (e.g. Westwood *et al.*, 2005), there are only a handful of papers which have reflected on the impact of study quality in reviews that have included 'qualitative' research to address questions about intervention processes or people's perspectives and experiences (Attree and Milton, 2006; Campbell *et al.*, 2003; Noyes *et al.*, 2005; Popay *et al.*, 2003; Sandelowski and Barroso, 2002). Despite interest in driving up the quality of 'qualitative' research (Blaxter *et al.*, 1996; Spencer *et al.*, 2003), and a number of studies documenting the quality characteristics of samples of 'qualitative' studies (e.g. Borreani *et al.*, 2004; Boulton *et al.*, 1996), there is much uncertainty about what the 'fatal flaws' might be in this type of work (Dixon-Woods *et al.*, 2004b).

Chapter six described one approach to assessing the quality of 'qualitative' and other types of 'non-trial' studies. The first part of this chapter reports the outcome of using this approach to assess the quality of two different types of 'non-trial' studies which often, but not always, use 'qualitative' methods: 1) process evaluations, which are designed to examine and/or monitor the way an intervention is delivered and received; and 2) studies of health or social phenomena from people's own perspectives and experiences. The second, third and fourth parts of this chapter report a series of analyses that attempt to explore whether there is any relationship between a study's quality and the results of the synthesis that the study is included in. The underlying question of the analysis is whether and how study quality affects the results of syntheses about intervention processes and people's perspectives and experiences.

7.2 Methods

i) Sources of data

a) Studies

Three sets of studies were used:

1) Sixteen **process evaluations** collected for a review of the effectiveness and appropriateness of peer-delivered health promotion (Harden *et al.*, 1999b). The process evaluations were published between 1990 and 1998. Nine were conducted in the UK, six in the USA and one in Germany. Studies evaluated interventions in secondary schools, further education colleges and community settings. Health topics included sexual health, smoking, drugs, alcohol and healthy eating. The age of study participants ranged from 11 to 21 years.

2) Thirty-five **studies of young people's perspectives and experiences** collected for a series of three reviews on the barriers to, and facilitators of, mental health, physical activity and healthy eating (Harden *et al.* 2001b; Rees *et al.*, 2001; Shepherd *et al.*, 2001). All of these studies were conducted in the UK and published between 1990 and 2000. The age of study participants ranged from 11 to 24 years.

3) Thirteen **studies of children's perspectives and experiences** collected for two reviews on the barriers to, and facilitators of, physical activity and healthy eating amongst children (Brunton *et al.*, 2003; Thomas *et al.*, 2003). All of these studies were conducted in the UK and published between 1991 and 2002. The age of study participants ranged from four to 11 years.

Detailed data on the characteristics of these studies had been collected using standardised guidelines by pairs of experienced researchers. These researchers had initially worked independently and then came together to compare and agree a final set of data for each study. The guidelines contained between 54 and 77 questions on study characteristics (depending on which version of the guidelines were used) covering: the aims, rationale and background to the study; sampling, recruitment and sample characteristics; and data collection and data analysis methods. Both the original studies and the standardised information collected on their characteristics were used as sources of data for the analyses reported in this chapter.

b) Quality assessment

Each study had been quality assessed by pairs of experienced researchers using the approach and tool described in chapter 5. The specific criteria used to assess quality in each of the three sets of studies varied. This was because the approach and tool developed over time and different versions became available for new reviews. The criteria are listed below together with details of which criteria were applied to which studies (a = process evaluations; b = studies of young people's perspectives and experiences; and c = studies of children's perspectives and experiences).

- Explicit account of theoretical framework and/ or literature review [a, b]
- Aims and objectives were clearly reported [a, b, c]
- Adequate description of the context in which the research was carried out (including a rationale for why the study was undertaken) [a, b, c]

- Adequate description of the sample used and the methods for how the sample was identified and recruited [a, b, c]
- Adequate description of the methods used to collect data [a, b, c]
- Adequate description of the methods used to analyse data [a, b, c]
- Sufficient original data was provided to mediate between data and interpretation [a, b]
- Analysis of data undertaken by more than one researcher [a]
- Attempts made to establish the validity of data analysis [b]
- Sufficient attempts to establish the reliability of data collection tools [c]
- Sufficient attempts to establish the validity of data collection tools [c]
- Sufficient attempts to establish the reliability of the data analysis methods [c]
- Sufficient attempts to establish the validity of data analysis methods [c]
- Appropriate data collection methods used for helping people to express their views [c]
- Appropriate methods used for ensuring data analysis grounded in people's views [c]
- Appropriate active involvement of representatives from the population being studied in the design and conduct of the study [c]

In short, the process evaluations and studies of young people's perspectives and experiences had been assessed according to how well their methods had been reported and how rigorous these methods were. Studies of children's perspectives and experiences had also been assessed according to how appropriate study methods were for answering the review question (i.e. were the methods suitable for studying children's perspectives and experiences?). These studies were assigned a 'weight of evidence' rating (low medium, or high) by reviewers. (For a description of the development of 'weight of evidence' see chapter 6.) Reviewers were asked

'What weight of evidence would you give this study in terms of whether its findings are really rooted in the perspectives of the people studied?' and, to make a judgement, reviewers were asked to consider a) the quality of the reporting on the methods used in the study; b) the rigour of the methods; and c) the appropriateness of methods for studying people's perspectives and experiences.

c) Synthesis

The findings of the studies had been synthesised in six systematic reviews: process evaluations were synthesised in a review of peer-delivered health promotion; studies of young people's perspectives were synthesised in three reviews on the barriers to, and facilitators of, mental health, physical activity and healthy eating; and studies of children's perspectives and experiences were synthesised in two reviews on the barriers to, and facilitators of, physical activity and healthy eating. Methods for synthesising findings varied across the reviews. There were three main types of synthesis conducted:

1) In the review of peer-delivered health promotion the synthesis of the findings of the 16 process evaluations produced a narrative structured according to the type of processes addressed by the study such as the accessibility and acceptability of interventions, or factors affecting intervention delivery or implementation. The question driving the synthesis was whether or not peer-delivered health promotion is appropriate for young people.

2) In the reviews on the barriers to, and facilitators of, young people's health, the synthesis of the findings of studies of young people's views produced lists of 'barriers and facilitators' associated with mental health, physical activity or healthy eating. 'Barriers' were defined as those factors that stopped young people from feeling bad, taking part in physical activity, or eating healthily or make it less likely

that young people will feel good, be physically active, or eat healthily. 'Facilitators' were defined as those factors that helped young people to feel good or cope with feeling bad, take part in physical activity, or eat healthily. In preparation for the synthesis reviewers had i) extracted and summarised the main or key findings of each study as highlighted by study authors in, for example, the study summary or conclusions; ii) added findings not highlighted by authors as their main or key findings; and iii) classified all findings according to whether they revealed a barrier or a facilitator (or a 'perception/meaning') (figure 7.1).

Figure 7.1: Illustration of extracted and summarised findings from one study in a review on young people and physical activity.

Study	**Key findings reported by authors**	**Reviewers' conclusions on young people's views**
Mitchell (1997)	*Barriers to participation: conflicting interests/lack of time; lack of motivation *Low participation rate in organized sports *Teenage magazines play a central role in young women's lives *Feasible and acceptable to promote physical activity in teenage magazines	***Perceptions of/meaning of physical activity*** *Preference for cycling, swimming, aerobics rather than organised team sports *Feeling fit, toned/getting a better figure; maintaining health, acquiring new skills and building confidence are the perceived benefits of exercise. *Young women noted that physical activity does not fit with usual content of magazines ("girlie stuff"). ***Barriers*** *Agree with authors ***Facilitators*** *Using magazines to promote physical activity (the young women had a preference for articles about readers engaging in sport rather than specific instructions for exercise)

The synthesis was completed in two stages. Firstly, all the barriers and facilitators from each individual study were listed together and their number was reduced to take account of those that had been identified by more than one study. For example, in the young people and physical activity review the 16 included studies yielded a total of 80 barriers and facilitators. These were collapsed into 40 separate barriers and facilitators. Secondly, the list of barriers and facilitators was organised into four headings to reflect the domains of young people's lives in which particular barriers and facilitators operated: 'family and friends' (e.g. parental or peer support);

'self' (e.g. personal resources such as attitudes or knowledge); 'practical and material resources' (e.g. money, time); and 'the school' (e.g. relationships with teachers, school canteen) (Figure 7.2).

Figure 7.2: Illustration of synthesis product from a review on the barriers to, and facilitators of, young people and physical activity.

Practical and material resources	
Barriers	**Facilitators**
*Lack of time (*Y5*, *Y7*, *Y10*, *Y11*, *Y13*) *Lack of money (*Y3*, *Y7*, *Y11*) *Provision of activities which are associated with childhood or primary school, are highly structured, or organised by adults (for young women) (*Y3*)	*Creation of more cycle lanes (*Y1*) *Make activities more affordable (*Y11*) *Increasing access to clubs for young people to dance (*Y11*) *Single sex physical activities at youth clubs with mixed sex (non-physical) activities afterwards (*Y11*) *Provision of more acceptable forms of physical activity such as aerobics (*Y14*) *More consensus about desirable health behaviour (*Y5*)

3) During the conduct of the reviews on the barriers to, and facilitators of, healthy eating and physical activity amongst children, methods for synthesis underwent some development. In the children and physical activity review, a list of barriers and facilitators was still the main synthesis product. The barriers and facilitators were grouped according to the underlying themes they suggested: 'preferences, priorities and valued aspects of physical activity' (e.g. not enjoying sport or exercise; playing sport is a way of forgetting troubles); 'family life and parental support' (e.g. practical support from parents); and 'restricted/greater access to opportunities' (e.g. lack of transport to get to facilities'). In the children and healthy eating review, the main synthesis product was a set of themes rather than a list of barriers and facilitators. There were several reasons for this but a significant factor associated with the change was the fact that the findings of the studies included in the children and

healthy eating review could not be read easily as straightforward expressions of barriers or facilitators. Rather they described children's beliefs and attitudes and/or the meanings and place of food, eating, and health in children's lives. The team had to find a way of synthesising these kinds of findings or they would have been left with an empty synthesis. Rather than try to identify barriers and facilitators in study findings and then organise these into themes, the team decided to conduct a full thematic analysis of the findings from the start. All study findings were entered into a software package for aiding 'qualitative' analysis and line by line coding was undertaken. Codes were grouped, deleted and collapsed into a smaller number so that the findings of the original studies were distilled down into their essential features and then combined into a whole via a listing of themes.

This process resulted in thirteen 'descriptive themes' (e.g. 'good and bad foods'; 'food preferences'). The team then moved to a higher level of abstraction by thinking about what children's perspectives and experiences (as represented by the descriptive themes) were suggesting for what might help them, and what might stop them, eating fruit and vegetables. Six 'analytic themes' emerged from this process (e.g. 'children prioritise taste over health for choosing food'; 'fruit and vegetables have different meanings for children') and these were used to generate nine implications for intervention development (e.g. 'reduce emphasis on health messages'; 'promote fruit and vegetables in different ways').

ii) Analysis

To describe the quality of studies the number of studies meeting each quality criterion was calculated. The relationship between study quality and synthesis results was explored in different ways according to the three types of synthesis described above. As already noted earlier, none of the studies had been excluded

from the syntheses on the grounds of quality. This provided an opportunity to study whether there was any relationship between the quality of studies and the role or contribution of their findings in the syntheses they were included in.

a) Study quality and synthesis results in a process synthesis.

Examining whether there was any relationship between the quality of process evaluations and the synthesis results was complicated by the fact that the studies were addressing a range of processes and had multiple findings. The analysis avoided this complication by focusing on claims made by study authors which answered the question of the review on the appropriateness or otherwise of peer-delivered health promotion for young people. The analysis was driven by the assumption that higher quality studies would be more likely than low quality studies to minimise the introduction of bias and error. If there were differences in findings about the appropriateness of peer-delivered health promotion, the findings from higher quality studies would be seen as a more reliable approximation of the true answer. Studies were divided into those who concluded that young people had largely positive views and experiences, largely negative views and experiences, or mixed views and experiences of peer-delivered health promotion. The quality of studies across these categories was examined.

b) Study quality and synthesis results in syntheses of young people's views.

The question in the review of peer-delivered health promotion required an answer in the form of a 'yes' or a 'no'. In contrast, the review questions for the series of reviews that included studies of people's perspectives and experiences required answers in the form of a list (e.g. What are the barriers to, and facilitators of, healthy eating amongst young people?). This type of synthesis is like a survey or a summary of the findings of individual studies as it brings together and describes in one place all of the barriers and facilitators located within individual studies.

Although this synthesis does pool findings that are the same, it does not aim to pool all findings to produce one overall finding like a statistical meta-analysis. Rather, its goal is to produce a 'jigsaw' or 'mosaic' to reveal a more complete picture of the phenomenon under study. This type of synthesis does transform findings by ordering and grouping them thematically, but it is not what some have called an 'interpretive synthesis' which aims to transform the findings of individual studies into higher order concepts or theories (Dixon-Woods *et al.*, 2005; Gough and Elbourne, 2002; Sandelowski and Barroso, 2003a). Within this type of synthesis, it was hypothesised that higher quality studies would contribute more pieces of the 'jigsaw'.

To examine whether higher quality studies did contribute more to the syntheses they were included in, a measure of 'synthesis contribution' was calculated for each study. This measure was informed by the work of Sandelowski and Barroso (2003b, p232) who calculated "intensity effect sizes" for each of the 45 studies included in their meta-summary of HIV-positive women's experiences of motherhood. In this meta-summary a total of 93 thematic statements, or abstracted findings, were distilled from the 45 studies. Each study's 'intensity effect size', defined as "the concentration of findings" in any one study, was calculated by dividing the number of abstracted findings produced by the study by the total number of abstracted findings in the synthesis overall. In the present analysis, each study's synthesis contribution, or 'intensity effect size', was calculated by dividing the number of barriers and facilitators identified by that study and dividing this by the total number of barriers and facilitators identified in the synthesis overall. For example, in the mental health review, the study by Aggleton *et al.* (1995) revealed 17 barriers and facilitators. Because a total of 82 different barriers and facilitators were identified by the review overall, Aggleton *et al.* (1995) had a synthesis contribution score of 17/82 or 21%.

A unique synthesis contribution measure was also calculated in order to examine what would be lost to a synthesis if a particular study had been excluded. For example, the study by Aggleton *et al.* (1995) revealed 17 barriers and facilitators and 11 of these were not found by any other study. This study's unique contribution score was therefore 11/82 or 13%. Each study's synthesis contribution and unique synthesis contribution scores were plotted against the number of quality criteria the study met. Statistical analysis was used to help interpret the plots, together with an analysis to compare and contrast the characteristics of high and low contributing studies.

c) Study quality and synthesis results in syntheses of children's views.

A measure of synthesis contribution was also used to examine the role of high and low quality studies in two syntheses of children's perspectives and experiences on physical activity and healthy eating. Again the goal of these syntheses was to produce a 'jigsaw' or 'mosaic' to reveal a more complete picture of the phenomenon under study than could be revealed by any one study alone. Both syntheses transformed findings by ordering and grouping them thematically, but the children and healthy eating review went one step further towards an 'interpretive synthesis' because it transformed the findings of individual studies into higher order concepts or theories in the form of analytical themes. Within both of these syntheses, it was hypothesised that higher quality studies would contribute more to the synthesis. In the physical activity review the assumption was that higher quality studies would contribute a larger number of barriers and facilitators. In the children and healthy eating review it was hypothesised that higher quality studies would contribute to a larger number of the descriptive and analytical themes.

For the studies in the physical activity review the synthesis contribution score was calculated by dividing the number of barriers and facilitators revealed by a particular

study by the total number of barriers and facilitators identified by the synthesis (n=39). (A unique synthesis score was also calculated by dividing the number of unique barriers and facilitators in a study by the total number of barriers and facilitators found overall.) For the studies in the children and healthy eating review the synthesis contribution score was calculated by dividing the number of descriptive themes a particular study contributed to by the overall number of descriptive themes to emerge from the synthesis (n=13). (A unique synthesis score was also calculated by dividing the number of descriptive themes not found by any other study by the total number of descriptive themes to emerge overall.)

As described earlier, in these reviews the quality of studies had been assessed according to the appropriateness of study methods for finding out about children's perspectives and experiences, as well as reporting quality and rigour of methods. These assessments reflected how confident reviewers were that study findings were rooted in children's perspectives and experiences. Each study's synthesis contribution and unique synthesis contribution scores were mapped against the number of quality criteria the study met. Again, statistical analysis was used to help interpret the plots, together with an analysis to compare and contrast the features of high and low contributing studies.

7.3 Results

i) Overview of study characteristics and quality

There were 62 separate studies available for analysis across the three groups described above (two studies were common to two groups). Study reports were published in a variety of formats but nearly two thirds were published in journals (N=42). Charities, health promotion departments, independent research

organisations, or national authorities published a third of the study reports (N=18). The remainder were published in books (N=2).

There was a diversity of methods used in the studies with some studies using more than one method (table 7.1). It was difficult to characterise studies as 'qualitative' or 'quantitative'. Many study authors combined techniques usually associated with one tradition or the other. For example, some studies had collected people's views in their own words and then translated these into numbers and used statistics to analyse them. In addition, some studies that had included both open-ended and fixed response items in their self-completion questionnaires did not always analyse answers to these open-ended questions.

Table 7.1: Methods of data collection and analysis across studies (N=62)

	N	**%**
Methods of data collection		
Fixed response self-completion questionnaire	18	29
Fixed and open response self-completion questionnaire	10	16
Open response self-completion questionnaire	1	2
Interviews and/or focus groups	30	48
Interviews and/or focus groups combined with self-completion questionnaire	2	3
Observation	1	2
Total	*62*	*100*
Methods of data analysis		
Descriptive and/or inferential statistics	24	39
'Qualitative' data analysis	27	43
Combination	11	18
Total	*62*	*100*

The methodological quality of the studies was variable (table 7.2). Whilst the majority of studies clearly reported their aims and objectives, between a third and

one half of studies failed to meet other basic criteria for reporting study context and other aspects of study methods.

Table 7.2: The quality of 'qualitative' and other types of studies as judged in six systematic reviews in health promotion and public health

	Process evaluations (N=16)	**Young people's views (N=35)**	**Children's views (N=13)**
An explicit theoretical framework and/or literature review	7 (47%)	16 (46%)	-
Aims and objectives clearly reported	11 (73%)	32 (91%)	11 (85%)
Adequate description of the context of the research	10 (67%)	31 (88%)	8 (62%)
Adequate description of the sample and how it was recruited	8 (53%)	17 (49%)	6 (46%)
Adequate description of data collection and analysis methods	7 (47%)	22 (63%)	-
Adequate description of data collection methods	-	-	11 (85%)
Adequate description of data analysis methods	-	-	3 (23%)
Sufficient original data to mediate between data and interpretation.	10 (67%)	21 (60%)	-
Analysis of data undertaken by more than one researcher.	3 (20%)	-	-
Attempts made to establish the reliability and validity of data analysis	-	6 (17%)	-
Sufficient attempts to establish reliability of data collection tools.	-	-	8 (62%)
Sufficient attempts to establish validity of data collection tools.	-	-	10 (77%)
Sufficient attempts to establish reliability of the data analysis methods.	-	-	6 (46%)
Sufficient attempts to establish validity of data analysis methods.	-	-	2 (15%)
Appropriate data collection methods used for helping people to express their views.	-	-	12 (92%)
Appropriate methods used for ensuring data analysis grounded in people's views.	-	-	5 (38%)
Appropriate active involvement of study population in design or conduct of the study	-	-	4 (31%)

Reporting quality was most problematic with respect to reporting data analysis methods adequately and presenting an adequate description of the sample and how

it was recruited. Like the scant information provided on some aspects of study methods, some study authors also failed to undertake, or provide detail on, attempts to increase the rigour of their data collection tools and data analysis methods. Compared to data collection, attempts to increase the rigour of data analysis were much less frequent.

The thirteen studies focused on children were also assessed in terms of the appropriateness of their methods for finding out about children's perspectives and experiences. All but one study was judged to have used appropriate data collection tools, but only five were judged to have used appropriate data analysis methods. The latter figure does not mean that the other studies did not use appropriate methods. In most cases reviewers could not tell because of the minimal detail provided on how data were analysed. Only four studies were judged to have actively involved representatives from the population being studied in the design and conduct of the study.

ii) Study quality and a synthesis of process evaluations

In the review of peer-delivered health promotion, 11 of the 16 included process evaluations had assessed, and drawn conclusions about, how acceptable young people found peer-delivered health promotion to be. Of the 11 studies, six found only positive appraisals, whilst five found a mixture of positive and negative appraisals. Studies meeting a higher number of quality criteria were more likely to find mixed appraisals than those meeting a lower number of quality criteria (table 7.3). Three of the 11 process evaluations examining acceptability met four or more of the quality criteria and one of these found only positive appraisals (or 33 per cent) whilst the other two found a mixed appraisal (or 67 per cent). Of the eight studies

that met three or less of the quality criteria, five found only positive findings (or 62 per cent) and three found a mixed appraisal (or 38 per cent).

Table 7.3: Findings according to the number of quality criteria met in process evaluations that drew conclusions on the acceptability of peer education (N=11)

	Process evaluations according to the number of quality criteria met						
	1	2	3	4	5	6	7
Mixed appraisal of peer education	Frankham (1993)	-	Newman *et al.* (1990) Strousse *et al.* (1990)	Fox *et al.* (1993)	-	-	Schonbach (1995)
Positive appraisal of peer education	-	Peers *et al.* (1993)	Fife Healthcare NHS Trust (1996) Guy and Banim (1990) Orme and Starkey (1999) Richie *et al.* (1990)	-	Chaiken (1990)	-	-
Negative appraisal of peer education	-	-	-	-	-	-	-

However, looking at individual studies, this pattern is not clear-cut. The one study that met all seven criteria (Schonbach, 1995) found a mixed appraisal. One study met five of the criteria and this study found only positive appraisals (Chaiken, 1990). One study met four of the criteria and found a mixed appraisal (Fox *et al.*, 1993). Six studies met three of the criteria and four of these found only positive appraisals (Fife Healthcare NHS trust, 1996; Guy and Banim, 1990; Orme and Starkey, 1999; Richie *et al.*, 1990) and two found mixed appraisals (Newman *et al.*, 1990; Strousse *et al.*, 1990). One study met two of the criteria and found only positive appraisals (Peers *et al.*, 1993) and one study met only one of the criteria and found a mixed appraisal (Frankham, 1993).

The process evaluations did not always stick to drawing conclusions about process. The conclusions of each of the process evaluations were examined in order to judge whether they were 'warranted' by the methods and findings of the research (i.e. did they make claims beyond what their methods and results would allow them to do?). This was plotted against the number of quality criteria that the study met (table 7.4). Overall, there were four process evaluations whose conclusions were not found to be warranted by their findings and methods (Fife Healthcare NHS Trust, 1996; Fox *et al.*, 1993; Frankham, 1993; Guy and Banim, 1990). Although the numbers are extremely small, none of these studies met more than four of the quality criteria. Interestingly, all these studies had made inappropriate conclusions about the impact of the intervention. This may reflect study authors being unclear about the limitations of process evaluations in drawing conclusions about the effects of interventions. This suggests that the seven criteria might be useful for helping to distinguish those studies with a thorough understanding of the strengths and limitations of the methods that they employ.

Table 7.4: Number of studies in which conclusions were judged to be 'warranted' by study methods and findings according to the number of quality criteria met.

Studies meeting.....	Studies for which conclusions judged to be 'warranted' by their methods and findings	Studies for which conclusions judged not to be 'warranted' by their methods and findings
7 of the quality criteria (N=2)	2	0
5 of the quality criteria (N=2)	2	0
4 of the quality criteria (N=2)	1	1
3 of the quality criteria (N=6)	4	2
2 of the quality criteria (N=2)	2	0
1 of the quality criteria (N=1)	0	1

(iii) Study quality and syntheses of young people's perspectives and experiences

a) Overview

When the quality of the 35 studies of young people's perspectives and experiences was mapped against their overall synthesis and unique synthesis contribution there is a suggestion of a positive relationship between study quality and contribution to synthesis but the relationship is far from clear cut (figure 7.3 and figure 7.4). Indeed, when the relationship between study quality and overall synthesis contribution was analysed statistically, the correlation co-efficient was positive but not statistically significant[1]. Comparing each quality criterion individually, to see whether some were better than others for predicting a study's synthesis contribution score, did not reveal any clear patterns either (e.g. was 'Are there attempts made to establish the reliability and/or validity of the data analysis?' any better at predicting a study's synthesis contribution score than 'Did the report explicitly and clearly state the aims of the study?')[2].

Some studies do follow the pattern one might expect in figure 7.3: there are low quality studies that have low synthesis contribution scores (e.g. P_1, P_9, P_8, P_{16}, H_5, M_4, M_{10}) and there are high quality studies that have high contribution scores (e.g. P_3, P_{11}, M_9). There is, however, a considerable number of studies which have high quality scores but low synthesis contribution scores (e.g. P_2, P_6, H_1, H_2) and three studies which are low quality but have a relatively high synthesis contribution score (H_4, M_6, $P1_2$,). The characteristics of these four groups of studies were compared and

---

[1] Pearson r = 0.319, p>0.05 (2-tailed test). It is likely that this result is overly influenced by the extreme data points P3, P11 and M9. When these cases are removed from the analysis r = 0.100, p = 0.58 (2-tailed test).

[2] R^2 = 0.301 F (7, 28) = 1.725, P>0.05

Figure 7.3: The relationship between the quality of 'qualitative' studies of young people's perspectives and experiences and their contribution to three syntheses on mental health, physical activity and healthy eating

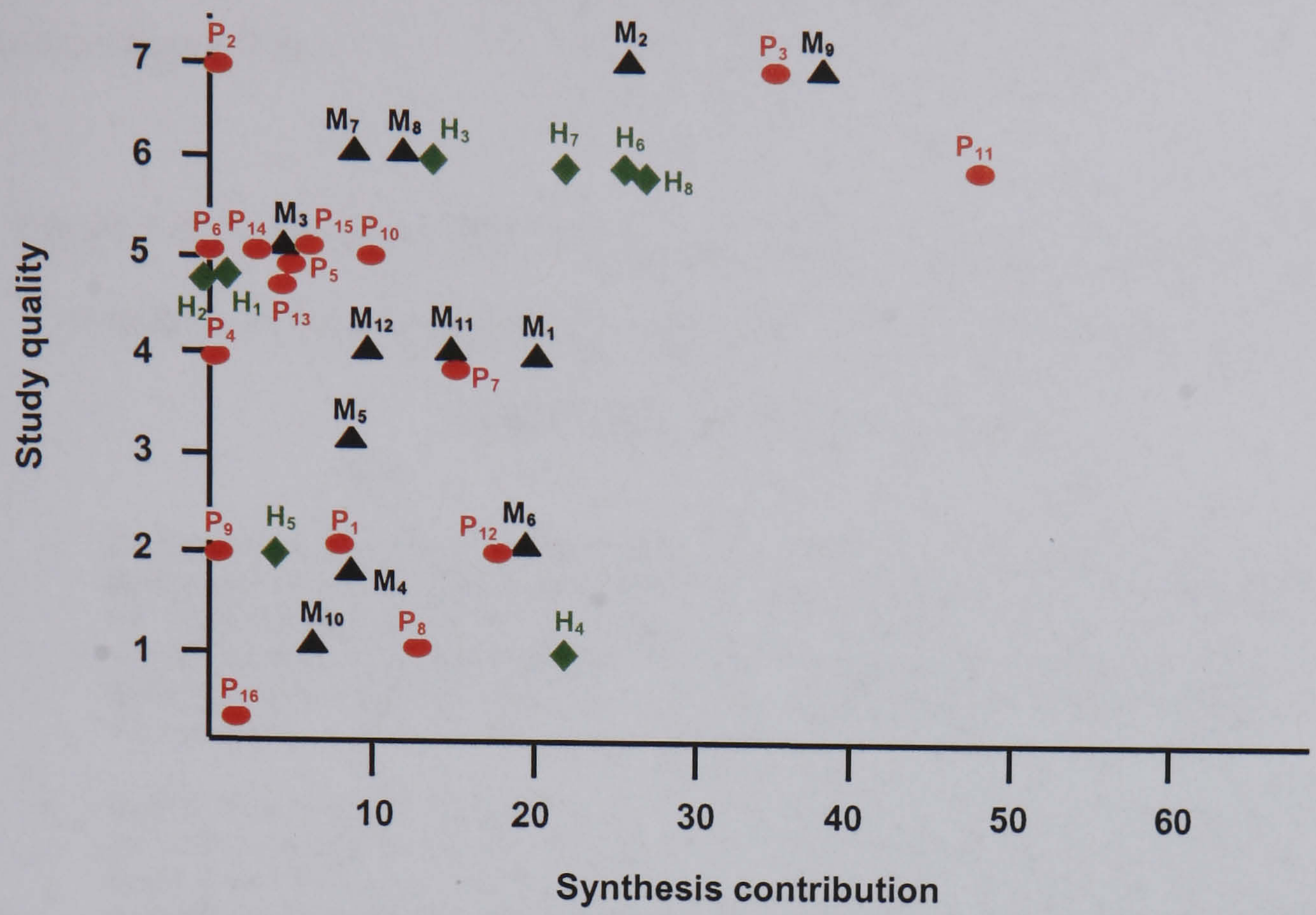
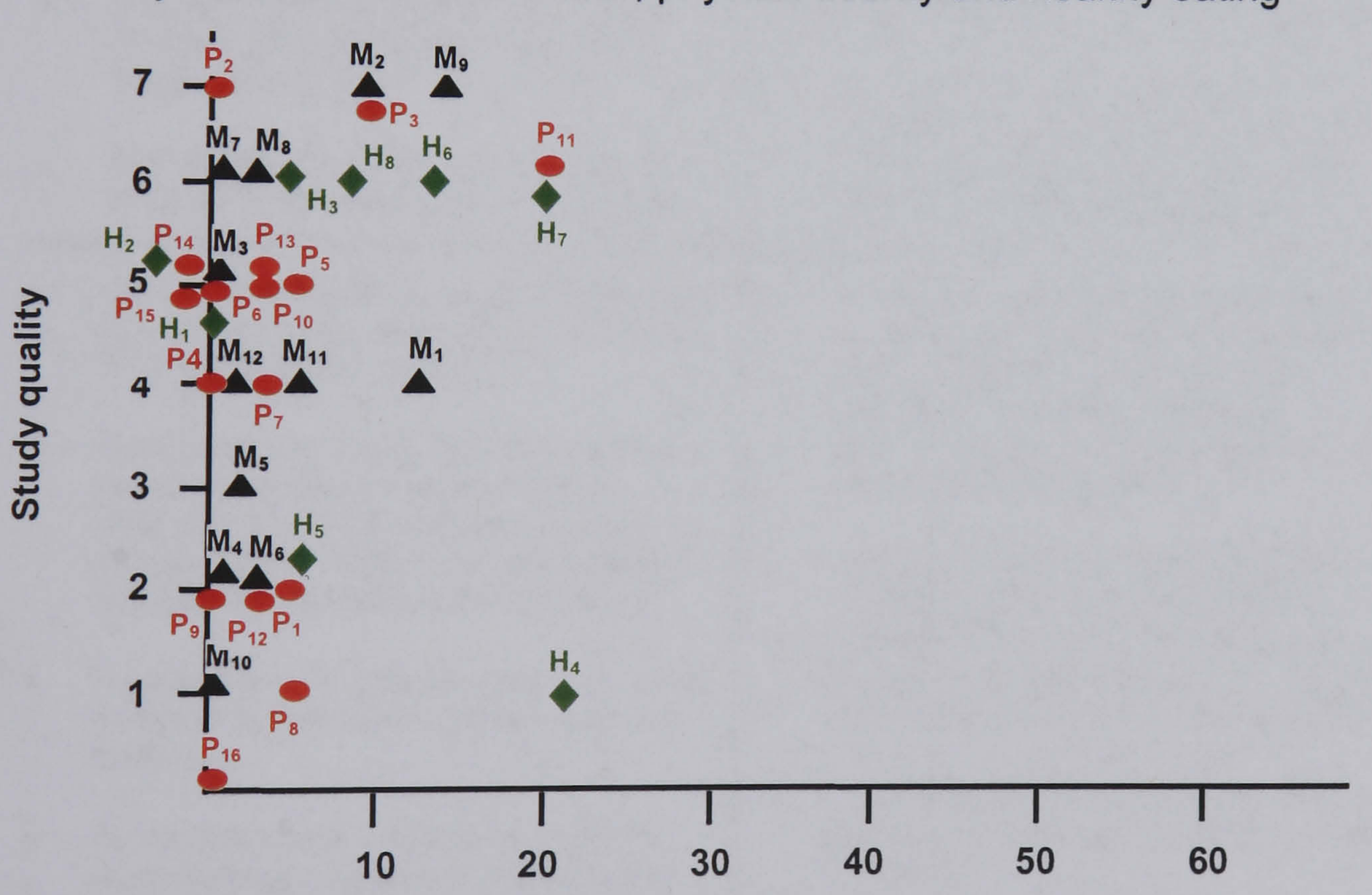


Figure 7.4: The relationship between the quality of 'qualitative' studies of young people's perspectives and experiences and their *unique* contribution to three syntheses on mental health, physical activity and healthy eating



Key for mental health studies

M_1 Aggleton et al. (1995)
M_2 Armstrong et al. (1998)
M_3 Balding et al. (1998)
M_4 Bowen (1997)
M_5 Derbyshire (1997)
M_6 Friedli and Sherzer (1996)
M_7 Gallagher et al. (1992)
M_8 Gallagher and Millar (1996)
M_9 Gordon and Grant (1997)
M_{10} Health Education Authority (1995)
M_{11} Scott-Porter (2000)
M_{12} Tolley et al. (1998)

Key for physical activity studies

P_1 Balding et al. (1997)
P_2 Birtwistle and Brodie (1991)
P_3 Coakley and White (1992)
P_4 Gentle et al. (1994)
P_5 Harris (1993)
P_6 Hopwood and Carrington (1994)
P_7 Kincey et al. (1993)
P_8 Mason (1995)
P_9 Miller (1993)
P_{10} Mitchell (1997)
P_{11} Mulvihill et al. (2000a)
P_{12} Orme (1991)
P_{13} Rogers et al. (1997)
P_{14} Sports Council Wales (1994a)
P_{15} Sports Council Wales (1994b)

Key for healthy eating studies

H_1 Dennison and Shepherd (1995)
H_2 Harris (1993)
H_3 McDougall (1998)
H_4 Miles and Eid (1997)
H_5 Roberts et al. (1999)
H_6 Ross (1995)
H_7 Watt and Sheiham (1996)
H_8 Watt and Sheiham (1997)

contrasted. The results of this comparison are reported in more detail below but a summary of the features that emerged to characterise each of the four groups is shown in figure 7.5.

Figure 7.5: Summary of study features according to quality and contribution to synthesis in three reviews about the promotion of young people's health

QUALITY	CONTRIBUTION TO SYNTHESIS: low	CONTRIBUTION TO SYNTHESIS: high
high	• Study aims are NOT a close match to the review aims in terms of one or more of the following: a) barriers and facilitators; b) young people's perspectives and experiences; and c) developing interventions. • Quality of reporting is good but studies do not always: a) report sufficient detail about the sample or sampling; b) attempt to establish the reliability and validity of data analysis; or c) report sufficient original data to support their analysis. • Studies often, but not always, use self-completion questionnaires with fixed response options • Reviewers found it difficult to interpret study findings as barriers or facilitators.	• Study aims are a close match to the review aims in terms of all of the following: a) barriers and facilitators; b) young people's perspectives and experiences; and c) developing interventions. • Quality of reporting is good and attempts are made to establish the reliability and validity of data analysis. Studies do not always report sufficient original data to support their analysis. • Studies often, but not always, use open-ended data collection techniques. • Study findings are relevant and detailed descriptions of barriers and facilitators, and display conceptual depth and explanatory power.
low	• Some studies do have aims and findings that are a close match to the review aims BUT some studies do NOT. • Quality of reporting is poor and there are no attempts made to establish the reliability and validity of data analysis. Studies do not always present sufficient original data to support their analysis. • A variety of methods are used (e.g. open ended or fixed response data collection methods). • Reviewers found it difficult to interpret study findings as barriers or facilitators AND/OR study findings lack detail and are limited in depth to lists of issues raised by the sample or proportions expressing particular views.	• Study aims and findings are a close match to the review aims in terms of all of the following: a) barriers and facilitators; b) young people's perspectives and experiences; and c) developing interventions. • Quality of reporting is poor and there are no attempts made to establish the reliability and validity of data analysis. Studies do not always present sufficient original data to support their analysis. • A variety of methods are used (e.g. open ended or fixed response data collection methods). • Study findings are relevant and detailed, but are limited in depth to lists of issues raised by the sample and/or proportions expressing particular views.

The results of the comparison suggest that quality is a secondary issue for making sense of why some studies contributed more than other studies to the syntheses

they were included in. Two other factors - a) the closeness of the match between the aims of the study and the aims of the review and b) the relevance, detail and depth of the study findings – appear to be key to explaining whether a particular study made a high or low contribution to the synthesis it was included in. The next part of this section illustrates these factors in more detail within each of the four groups of studies in figure 7.5. A final section examines in more detail the unique synthesis contributions of the studies.

<u>b) High quality studies with low synthesis contribution scores</u>

Five studies that met between five and seven of the quality criteria stand out in figure 7.3 as they did not contribute anything to the synthesis of barriers and facilitators:

- (P_2) **Birtwistle and Brodie (1991)** entitled 'Children's attitudes towards activity and perceptions of physical activity' aimed to examine young people's perceptions of school physical education, their feelings about being physically active, and their reasons for being inactive. Data were collected with a self-completion questionnaire with fixed response options.

- (H_1)**Dennison and Shepherd (1995)** entitled 'Adolescent food choice: an application of the theory of planned behaviour' aimed to increase understanding of the factors affecting food choice decisions. Data were collected with a self-completion questionnaire with fixed response options.

- (H_2) **Harris (1993)** entitled 'Young people's perceptions of health, fitness and exerciser' aimed to explore attitudes, beliefs and views with a particular focus on exploring difference by age and sex. Focus groups were used to collect data.

- (P_6) **Hopwood and Carrington (1994**) entitled 'Physical education and femininity' aimed to examine differences in the way young women and young men view and

experience sports and physical activity. Data were collected with a self-completion questionnaire with fixed and open response options.

- (P_{14}) **Sports Council Wales (study 1) (1994)** entitled 'A matter of fun and games: children's participation in sport' aimed to examine influences on young people's participation in sport and to document levels of involvement. Data were collected with a self-completion questionnaire with fixed response options.

At first sight these studies all appear to be highly relevant to reviews about young people's health. However, on closer inspection the aims and findings of these studies were not a very close match to the aims of the review they were included in. Unlike the reviews, these studies intended to contribute primarily to academic debate rather than to the development of interventions and were focused only indirectly on the barriers to, and facilitators of, young people's mental health, physical activity, and healthy eating. For example, although Birtwistle and Brodie (1991) suggest that the findings of their study might help teachers become more aware of the factors influencing attitudes to physical education, their main aim was to contribute to a body of knowledge about attitudes to physical activity. Other study authors in this group also emphasised the contribution studies would make to knowledge in the abstract. For example, Dennison and Shepherd (1995) aimed to test a theoretical model of behaviour to predict whether young people eat healthy foods and Hopwood and Carrington (1994) wanted to assess a recently made claim that sports participation is now more compatible with feminine self-image.

As might be expected given the aims of these studies, reviewers felt that studies did not present findings on barriers and facilitators. Although the study reported by Birtwistle and Brodie (1991) included 'examine young people's reasons for being inactive' as one of its aims, the findings of this study only covered young people's beliefs about the objectives of physical education at school (e.g. fitness, enjoyment);

their beliefs about the importance of physical education; and differences in attitudes according to age, sex and academic ability. Because none of these findings were linked to whether young people participated in physical activity or not, reviewers were reluctant to interpret these findings as barriers and facilitators to participation in physical activity. The studies by Hopwood and Carrington (1994) and Sports Council Wales (study 1) (1994) posed similar problems for reviewers. The findings presented by Hopwood and Carrington (1994) were the differences in attitudes towards physical activity between young men and young women but these were not linked to their participation levels. Similarly, although the Sports Council Wales (study 1)(1994) asked young people why they were inactive the study report only presented findings on which sports young people participate in and which ones they like the most and least. These problems were apparent in Harris (1993) too, but this study also appeared to have little to contribute because its findings presented young people's views on what the broader concept of health meant to them. Dennison and Shepherd (1995) did study young people's behaviour (the healthy eating choices they made) and used a theory from social psychology (the Theory of Planned Behaviour) to predict behaviour statistically from young people's attitudes. Reviewers, however, were reluctant to interpret these links as barriers and facilitators because young people themselves did not identify them.

c) Low quality studies with low synthesis contribution scores

Three studies which had been judged by reviewers as meeting either none, one or two of the quality criteria stand out as contributing very little to the synthesis in figure 7.3:

- (P_9) **Miller (1993)** entitled 'Femininity, physical activity and the curriculum' aimed to assess the extent of conflicts or ambiguities between perceptions of femininity amongst young women with very active lifestyles (sports and/or dance). Group interviews were used to collect data.

- (H_5) **Roberts *et al.* (1999)** aimed to study dieting behaviour as suggested by the title of the study report 'Dieting behaviour among 11-15-year-old girls in Merseyside and the Northwest of England'. Data were collected with self-completion questionnaires with fixed response options.

- (P_{16}) **Warburton (1998)** entitled 'Catch 'em young…Fit for life project' aimed to collect young people's views to inform the development of an intervention to promote physical activity. Focus groups were used to collect data.

Like the high quality studies which contributed nothing to the synthesis, the studies by Miller (1993) and Roberts *et al.* (1999) were not undertaken to inform the development of an intervention and their study aims and findings were not a close match to the aims and focus of the reviews they were included in (what helps and what stops young people taking part in physical activity/eating healthily). Miller (1993) undertook her study as she was interested in stereotypes of sports women as unfeminine. She found that conventional notions of femininity did not fit with young women's identities as physically active and that young women had to deploy strategies to make them fit (e.g. ensuring muscles are not too well developed). Reviewers argued that they could not use these findings to shed light on what young people see as barriers and facilitators to their participation in physical activity because it was the study author, rather than the young women, that had made the connection between femininity and participation in physical activity[3]. Roberts and colleagues argued that their study was needed because little is known about the dieting behaviour of young women in the UK and they suspected young women may

[3] Other reviewers found this argument to be too restrictive and it was abandoned in later reviews. This point raises interesting questions about a) the fit between explanations of behaviour offered by researchers and the participants themselves; and b) how different reviewers can interpret the findings of the same study in different ways.

be confusing dieting with healthy eating. Findings of this study were in the form of percentages of young women dieting and percentages of young women agreeing with various attitudinal statements (e.g. 66% thought dieting was good for their health). Reviewers cautiously inferred one barrier from this study (worries about weight may lead to unhealthy eating in the form of dieting).

Warburton (1998) was different from the other two studies as this study *was* undertaken to inform the development of an intervention and did appear to be a close match to the aims and focus of the physical activity review as its focus was on what could be done to make young people more physically active. This study, which used focus groups to collect data, did not meet any of the quality criteria as very little detail was provided on study methods. Few findings were presented (the whole study was written up in a two page journal article) and these contributed only one barrier to the synthesis ('activities on offer at school are only acceptable to those who are sporty') and one facilitator (young women found aerobics to be an exciting, interesting and inviting form of exercise).

d) High quality studies with high synthesis contribution scores

Three studies which had been judged by reviewers as meeting either six or seven of the quality criteria stand out in the top right hand corner of figure 7.3:

- (P_3) **Coakley and White (1992)**, entitled 'Making decisions: gender and sport participation among British adolescents' aimed to explore how young people make decisions about participating and non participating in sport. Interviews were used to collect data.

- (M_3) **Gordon and Grant (1997)**, entitled 'How we feel: an insight into the emotional world of teenagers' aimed to examine the emotional and mental health of young

people to inform interventions. Data were collected with a self-completion questionnaire with open-ended response options.

- (P_{11}) **Mulvihill *et al.* (2000a)** entitled 'Physical activity 'at our time'' aimed to explore a range of issues regarding young and physical activity including: barriers and motivations; preferences for different activities; and the role of parents and friends. Focus groups were used to collect data.

These three studies had several features in common. In addition to all of the studies being judged to be high quality by reviewers (two met all seven quality criteria but Mulvihill *et al.*, 2000a met six as they did not report any attempt to establish the reliability and validity of their data analysis), their study research questions, methods and findings were a close match to the focus of the reviews in which they were included in the following six ways:

i) All studies were funded by national or regional public organisations to inform specific efforts to intervene in people's lives. The study by Coakley and White (1992) aimed to explore how young people make decisions about participating and non participating in sport and was funded by the Greater London and South East Sports Council who wanted to know how best to campaign to get more young people involved in sport; the study by Gordon and Grant (1997) was funded by the Greater Glasgow Health Board to find out how to best address young people's emotional and mental health needs; and the study by Mulvihill *et al.* (2000a) aimed to explore a range of issues regarding young and physical activity and was funded by the Health Education Authority who wanted the findings to inform their 'Active for Life' campaign to get sedentary people to become more physically active.

ii) All study authors demonstrated a commitment to the value of studying the perspectives and experiences of young people as a route to generating new and important knowledge. Coakley and White (1992) reported that they wanted to gain a 'different' and 'more useful' understanding of why young people do or do not take part in sport by viewing young people as active agents who create their own sports lives; Gordon and Grant (1997) stated that they wanted to allow young people to express their feelings in their own words to provide a contrast to "stark statistics" or adult views which tend to see young people as "worry free or hedonistic"; and Mulvihill *et al.* (2000a) reported that their intended focus was upon the meanings young people attach to physical activity and their perceptions of barriers and facilitators and that they expected these to "vary from those conventionally identified in the literature" (p14).

iii) All three studies used methods of data collection which allowed young people to express themselves in their own words. Coakley and White (1992) collected data using interviews in a style which aimed to engage young people in a non-threatening conversation about their sports participation (e.g. 'why questions' were avoided); Gordon and Grant (1997) used a self completion questionnaire with open ended questions and writing activities to collect data (e.g. 'Three things which make me happy are...'; 'Imagine that you are writing your own diary and say exactly how you feel today'); Mulvihill *et al.* (2000a) used focus groups and 'ad hoc' interviews with a detailed topic and prompt guide provided for those collecting data.

iv) In all three studies the starting point for data analysis was the data representing young people's perspectives and experiences rather than any a priori coding scheme. Coakley and White (1992) transcribed what they thought were the key statements from young people and then interpreted these statements taking into account factors such as age, sex and social class; Gordon and Grant (1997) report

that they developed a coding scheme based on what the young people were saying and stated that they were careful not to impose their own framework on what young people said; and Mulvihill *et al.* (2000a) stated that the emphasis in their analysis was on identifying the range of views and perspectives offered by young people

v) In all three studies 'barriers and facilitators' could easily be seen in, and picked out of, study findings by reviewers. Coakley and White (1992) identified five themes within young people's accounts of their decision to either participate or not participate in sport (concerns about becoming adults; concerns about personal competence; constraints related to money, parents and the opposite sex; support and encouragement from significant others; past experiences of sport in school). Each theme was rich with factors that influenced whether young people participated or not. Gordon and Grant (1997) presented their findings in 11 sections (e.g. 'self-esteem – what makes young people value themselves' 'what young people do with their feelings'). Again, each section was rich with factors identified by young people as making them feel bad or bad and strategies young people used to help them feel better. Mulvihill *et al.* (2000a) presented their findings in seven sections and one of these focused solely on barriers to participation in physical activity ('barriers to involvement in physical activity') and another two focused solely on facilitators ('motivations for involvement in physical activity' and 'promoting physical activity').

vi) Findings in all three studies went beyond listing barriers and facilitators towards providing more in-depth or 'rich' descriptions that began to explain why something acted as barrier or facilitator. The best example of this is in the study by Coakley and White (1992) who found that decisions to participate or not participate in sport were bound up with young people's transitions to adulthood. Taking part in organised sports programmes or sports that they associated with childhood were seen as "a step backward in their development" (p25). Whilst for young men taking

part in sport tended to affirm their 'manhood', for young women, taking part in sports did not help them to 'negotiate the transition to womanhood'. Coakley and White (1992, p32) concluded that young people's decisions "to participate in sports was integrally tied to the way young people viewed themselves and their connections to the social world".

Some or all of the above features were also common to the four other high quality studies that had high synthesis contribution scores:

- (M_2) **Armstrong *et al.* (1998)**, entitled 'Listening to children', aimed to examine young people's views about, and understandings of mental health and to examine their ideas about help-seeking and relevant professionals. Interviews and focus groups were used to collect data.

- (H_6) **Ross (1995)**, entitled 'Do I really have to eat that?': a qualitative study of schoolchildren's food choices and preferences' aimed to explore young people attitudes and beliefs in order to better understand young people's food choices. Data were collected via focus groups.

- (H_7) **Watt and Sheiham (1996)** aimed to study dietary patterns as suggested by the tile of the study report 'Dietary patterns and changes in inner city adolescents'. The study also aimed to examine knowledge, skills, and beliefs about food and to assess factors influencing young people's ability to change their eating patterns. Data were collected via a self completion questionnaire with fixed response options

- (H_8) **Watt and Sheiham (1997)**, entitled 'Towards an understanding of young people's conceptualisation of food and eating' aimed to assess the meanings of food-associated concepts for young people and how these fit into their lives. Data were collected via individual interviews.

It is interesting to note that Watt and Sheiham (1996) was the only study in this group to use self-completion questionnaires with fixed response items. This feature is more likely to be associated with those high quality studies which did not contribute anything or very little to the synthesis. This illustrates that studies using this methodology can contribute substantially to syntheses of barriers and facilitators. However, these were less likely to produce the in-depth descriptions of young people's perspectives and experiences than studies using interviews or focus groups, yielding instead a list of factors cited most frequently by young people.

e) Low quality studies with high synthesis contribution scores

Three studies that were judged as low quality by reviewers stand out in the bottom right hand corner of figure 7.3:

- (M_6) **Friedli and Scherzer (1996)** entitled 'Positive steps: mental health and young people' aimed to examine how young people are affected by, and cope with, mental health problems. Data were collected via interviews.

- (H_4) **Miles and Eid (1997)** entitled 'The dietary habits of young people' aimed to elicit young people views on healthy eating (to feed them back to decision-makers) as well as to compare knowledge with behaviour. Data were collected via a self-completion questionnaire with fixed response options.

- (P_{12}) **Orme (1991)** entitled 'Adolescent girls and exercise: too much of a struggle?' aimed to examine the influences and constraints on young women's participation in physical activity. Data were collected via focus groups.

The aims of the studies by Friedli and Scherzer (1996) and Miles and Eid (1997) were a close match to the reviews they were included in. Friedli and Scherzer (1996) argued in the presentation of the background to their study that by listening

to young people and building on their feelings and ideas, we can make the most useful contribution to the promotion of mental health and Miles and Eid (1997) stated that "Our survey sought to compare young people's knowledge with their behaviour. We felt that their opinions were valid and should be listened to and fed back to decision makers" (p 46). In contrast, the aims of the study reported in Orme (1991), who did not link her study to interventions, were to explore why young women's interest and participation in physical activity starts to decline when they move to secondary school. However, the findings of all three studies were a close match to the reviews they were included in, although they did not display the richness or in-depth description of the higher quality studies which had high synthesis contribution scores. This was because findings were presented in very short journal articles (Orme, 1991) or because findings were presented in the form of proportions of young people raising particular issues (Friedli and Scherzer, 1996; Miles and Eid, 1997). As an example of the latter, Friedli and Scherzer (1996) presented their findings in five sections, two of which included 'barriers and facilitators ('what do young people worry about and how might they improve the quality of their lives?' and 'coping strategies'). These findings were presented in the form of proportions of young people raising particular issues (e.g. most young people find discussing problems with friends or relatives to be a useful strategy for coping with and preventing anxiety).

f) Unique contributions

Despite the fact that low quality studies sometimes contributed more than higher quality studies, it is interesting to note that not much would have been lost in the reviews on young people's health if studies meeting only one or two of the quality criteria had been excluded (figure 7.4 and table 7.5).

Table 7.5: Number of unique findings on barriers and facilitators found by studies in three reviews on young people's health

	No of unique barriers and facilitators (no. of studies)		
	High quality studies [a]	Medium quality studies [b]	Low quality studies [c]
Young people and mental health	29 (n=6)	17 (n=4)	2 (n=3)
Young people and physical activity	16 (n=9)	2 (n=2)	5 (n=5)
Young people and healthy eating	10 (n=6)	0 (n=0)	6 (n=2)

[a] Studies meeting five or more quality criteria; [b] Studies meeting three or four quality criteria; [c] Studies meeting one or two quality criteria

Studies judged to be low quality identified few barriers and facilitators that had not already been identified by other studies. For example, in the young people and mental health review, if studies meeting only one or two of the quality criteria had been excluded only two of the 82 barriers and facilitators identified overall in this review would have been lost. There were three studies that had been judged to meet only one or two of the quality criteria used in this review (Bowen, 1997; Derbyshire, 1996; Friedli and Sherzer, 1996), and only one of these had identified barriers and facilitators that had not been identified in any other study (Friedli and Sherzer, 1996). Of course, any loss of findings could mean an incomplete answer to the review question under study. Friedli and Sherzer (1996), who examined how young people are affected by, and cope with, mental health problems, found that increased employment opportunities and keeping busy were things that young people said stopped them feeling bad. No other studies had identified these issues and so they would be lost to the review. (Other studies had, however, identified similar issues such as unemployment and choosing and finding a job). Nonetheless, it could be argued that these findings are only useful if they have been produced by good quality research.

(iv) *Study quality and syntheses of children's perspectives and experiences*

a) Overview

When the quality of the 13 studies of children's perspectives and experiences was mapped against their overall synthesis contribution and unique synthesis contribution, there is a suggestion of a positive relationship between study quality and contribution to synthesis. Like the analysis of reviews focused on young people, the relationship is, however, not a straightforward one (figures 7.6 and figure 7.7). Again, when the relationship between study quality and overall synthesis contribution was analysed statistically, the correlation co-efficient was positive but not statistically significant[4].

Once more there are four categories of studies, two of which follow the predicted pattern: i) high quality studies which have high synthesis contribution scores (H_2, H_6, P_4) and ii) low quality studies which have low synthesis contribution scores (H_7, H_8, P_1, P_2); and another two which do not follow the predicted pattern: iii) high quality studies which low synthesis contribution scores (H_1, H_4) and iv) low quality studies which have high contribution scores (P_3,, H_3). These four groups of studies were compared and contrasted to identify a set of features to characterise each group.

Unlike the previous analysis of studies from the reviews focused on young people, this comparison was able to use the additional quality criteria that these studies had been assessed against: whether children had been involved in the design and/or conduct of the study; whether methods of data collection were appropriate for studying children's views; and whether methods were appropriate for grounding the analysis in children's perspectives. The features that emerged as characterising

[4] Pearson r = 0.317, p>0.05 (2-tailed test)

Figure 7.6: The relationship between the quality of 'qualitative' studies of children's perspectives and experiences and their contribution to two syntheses on physical activity and healthy eating

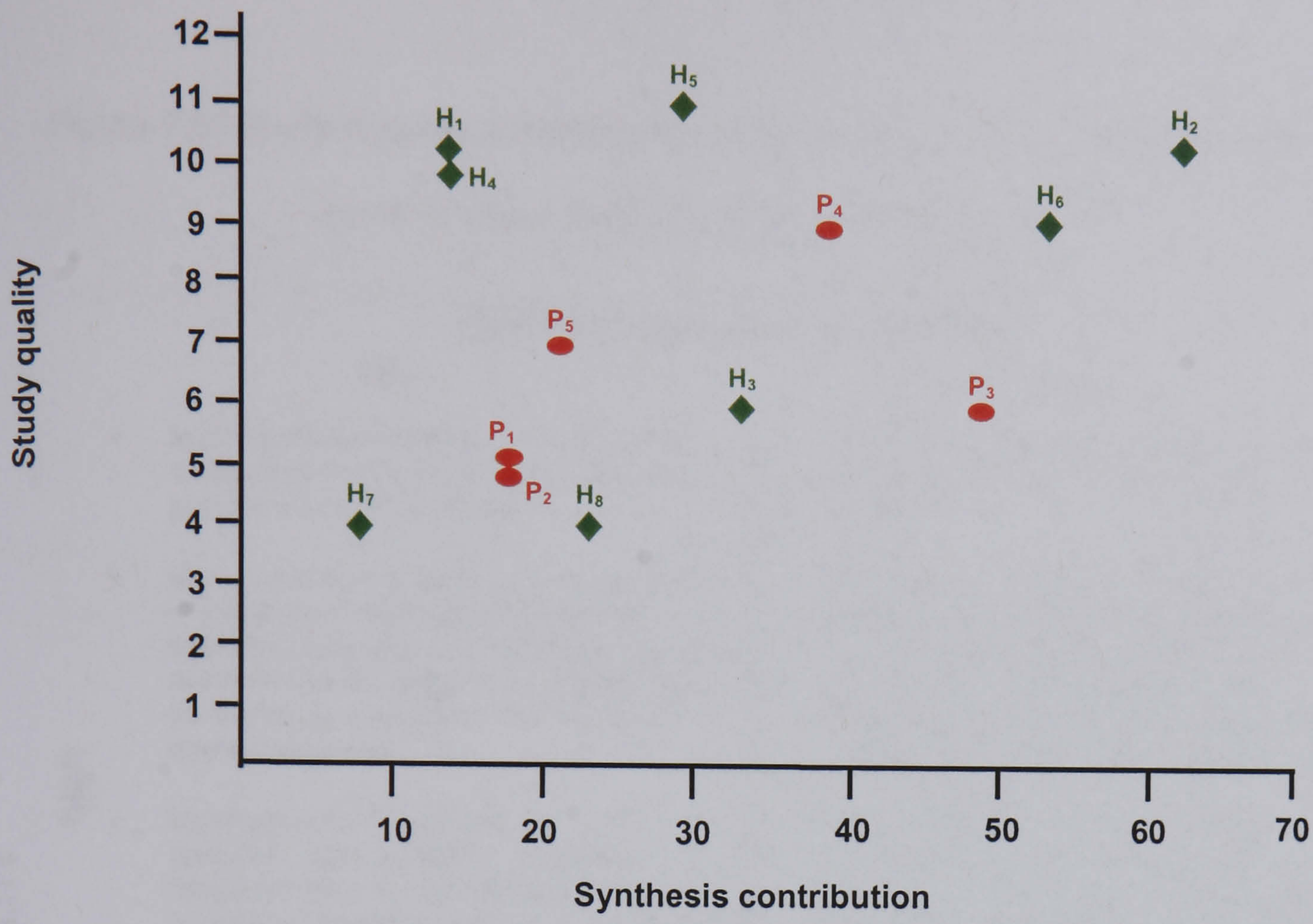
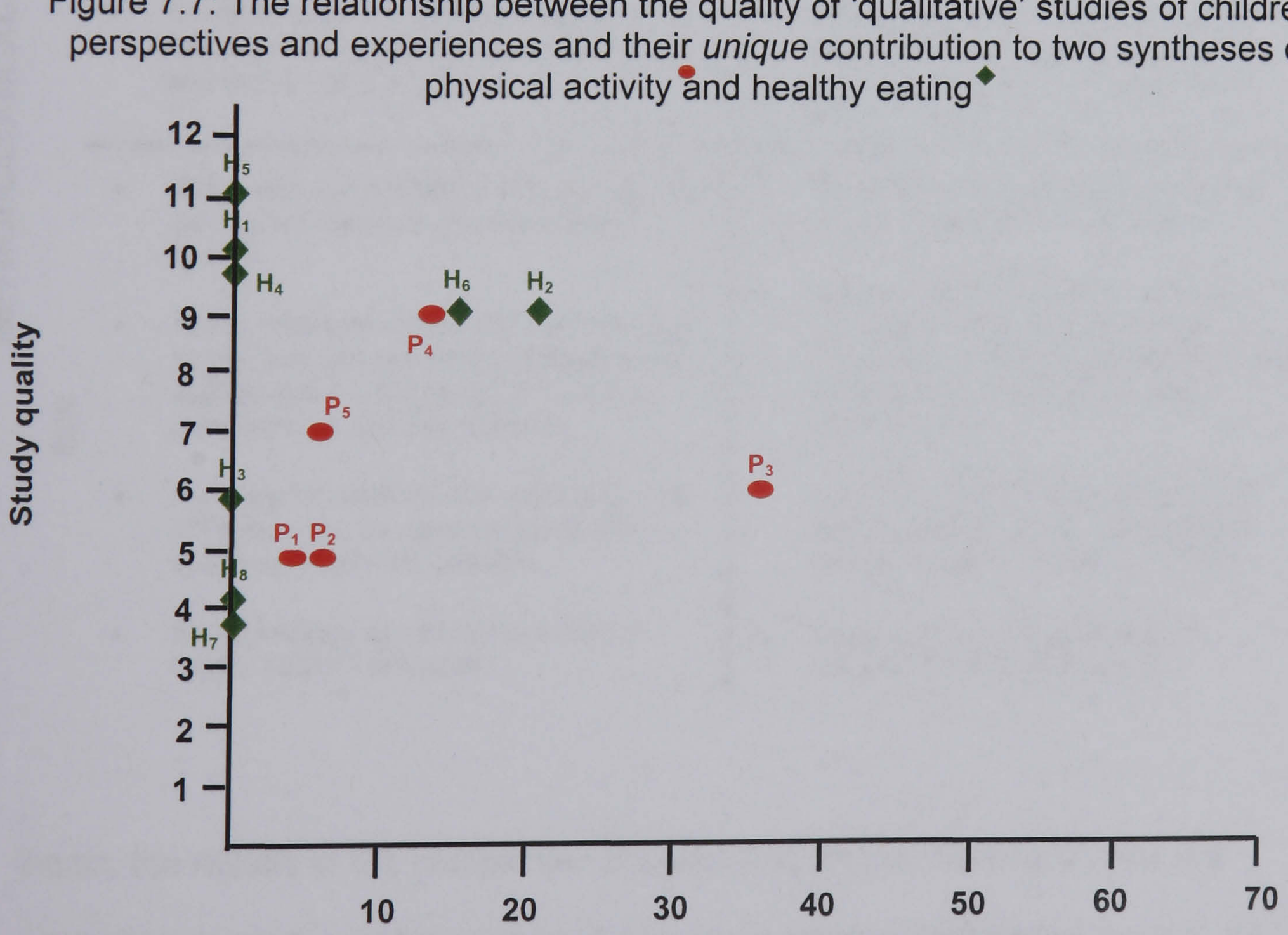


Figure 7.7: The relationship between the quality of 'qualitative' studies of children's perspectives and experiences and their *unique* contribution to two syntheses on physical activity and healthy eating



Key to healthy eating studies

H_1 Baxter et al. (2000)
H_2 Dixey et al. (2001)
H_3 Edwards and Hartwell (2002)
H_4 Gibson et al. (1998)
H_5 Hart et al. (2002)
H_6 Mauthner et al. (1993)
H_7 Neale et al. (1998)
H_8 Tilston et al. (1991)

Key to physical activity studies

P_1 Burrows et al. (1999)
P_2 Davies and Jones (1996)
P_3 Mason (1995)
P_4 Mulvihill et al. (2000b)
P_5 Tuxworth (1997)

each of the four groups of studies were remarkably similar to those that emerged in the analysis of the reviews focused on young people (figure 7.8).

Figure 7.8: Study features according to quality and contribution to synthesis in two reviews about the promotion of children's health

METHODOLOGICAL QUALITY	CONTRIBUTION TO SYNTHESIS: low	CONTRIBUTION TO SYNTHESIS: high
high	• In comparison to the aims of the review study aims, methods, and findings have a precise and narrow scope. • Methods are well reported with strategies in place for increasing rigour in data collection and analysis, but study methods are not always judged to be appropriate for the study of children's perspectives and experiences. • Most variables or concepts are pre-specified, data collection uses fixed response options, and statistical analysis is used to identify and explain patterns in the data. • Study findings are precise and narrow in scope, but may display conceptual depth and explanatory power.	• Study aims, methods, and findings are a close match to the aims of the review. • Methods are well reported and are judged to be highly appropriate for the study of children's perspectives and experiences, but studies are not always judged to have used strategies to increase rigour in data analysis. • Few variables or concepts are pre-specified, data collection is open-ended, and thematic analysis is used to identify and explain patterns in the data. • Study findings are detailed, cover a wide scope, and may display conceptual depth and explanatory power.
low	• Study aims and findings may or may not be a close match to the aims of the review. • Study methods are poorly reported, lack rigour, and are not always judged to be appropriate for the study of children's perspectives and experiences. • A variety of methods are used (e.g. data collection can be open ended and/or use fixed response options). • Study findings are sketchy, limited in depth, and/or relevance.	• Study aims and findings are a close match to the aims of the review. • Study methods are poorly reported, lack rigour, and are not always judged to be appropriate for the study of children's perspectives and experiences. • A variety of methods are used (e.g. data collection can be open ended and/or use fixed response options). • Study findings are detailed and relevant but limited in depth.

Again, the results of the comparison suggest that methodological quality is a secondary issue for making sense of why some studies contributed more to the syntheses they were included in than other studies. Two other factors - a) the relationship between the focus of the study and the focus of the review and b) the

scope, relevance, detail and depth of the study findings – appear to be key to explaining whether a particular study made a high or low contribution to the synthesis it was included in. These factors are illustrated in the next part of this section, which compares and contrasts the four groups of studies in figure 7.8 in more detail. A final section examines in more detail the unique synthesis contributions of the studies.

b) High quality studies with low synthesis contribution

Two high quality studies stand out in figure 7.6 as they contributed relatively little to the healthy eating synthesis they were included in:

- (H_1) **Baxter *et al***. (2000) entitled 'Children's perceptions of and preferences for vegetables in the West of Scotland' asked children to rate their preference for eight commonly consumed vegetables and explored why particular vegetables were liked and disliked by examining the sensory and 'eating occasion' characteristics the children assigned to different vegetables. Differences in perceptions and preferences according to socio-demographic variables were also examined. Interviews and then rating scales were used to collect data.

- (H_4) **Gibson *et al***. (1998) entitled 'Fruit and vegetable consumption, nutritional knowledge and beliefs in mothers and children' aimed to examine the impact of children's and mother's beliefs and knowledge on children's intake of fruit, vegetables and confectionary. Data were collected via interviews with fixed response options, self-completion questionnaires and food diaries.

In contrast to other studies in the children and healthy eating synthesis, these studies had a relatively specific and narrow focus. The studies were relevant to the review question, but their focus was quite specific (i.e. food likes and dislikes) and influencing factors were pre-specified in advance (e.g. sensory properties of

vegetables, socio-demographic variables, mother's beliefs). Relevant study findings were also specific and narrow. Both studies contributed to two of the thirteen descriptive themes to emerge from the healthy eating synthesis ('food preferences' and 'perception of health benefits'). Baxter *et al.* (2000) found that children preferred brightly coloured, small, soft, juicy and sweet vegetables. The vegetables that were disliked by children were large, hard and leafy and these were also the ones that the children associated with health benefits. Gibson *et al.* (1998) found that children rated taste as the most important factor in choosing food for themselves (as opposed to, for example, health benefits or eating the same foods as others), liked fruit almost as much as they liked sweets, and disliked vegetables.

A mis-match between study aims (and findings) and the aims of the review, rather than quality, therefore explains the low synthesis contribution of these studies. In contrast to the reviews about young people, this mis-match was not so much about whether or not a study focused on barriers and facilitators and the development of interventions - as noted earlier in the methods section, the syntheses in the reviews focused on children were less concerned over whether studies expressed their findings in terms of 'barriers and facilitators' – but because study aims were cast in relatively specific and narrow terms compared to the aims of the review (e.g. 'which vegetables do children prefer and why?' as opposed to 'what are children's perspectives on, and experiences of, eating fruit and vegetables?'). This meant that study findings were relatively specific and narrow in scope, and made only a small contribution to the synthesis (e.g. 'children prefer small, sweet and juicy vegetables' as opposed to 'children described the following five factors which influenced their intake of fruit and vegetables…'). Although findings were specific and narrow in scope, unlike the lower quality studies they did not lack conceptual depth and explanatory power. For example, Gibson *et al.* (1998) tested several variables for their ability to explain the variance in children's fruit and vegetable intake and found

that fruit intake was predicted by one set of factors (the beliefs and attitudes of mothers) and vegetable intake by another (children's preferences).

It is also worth noting that the methods used in these studies – pre-specified variables for analysis and data collection methods with fixed response options - though rigorous and well reported, were not always judged to be appropriate for studying children's perspectives and experiences. Although both studies were judged to have used appropriate data collection methods for helping children express their views, neither of the studies actively involved children in the design or conduct of the study. In addition, reviewers did not consider Gibson *et al.* (1998) to have used appropriate methods for ensuring the analysis was grounded in children's perspectives and experiences because the variables for analysis were not created with children's input.

<u>c) Low quality studies with low synthesis contribution</u>

One low quality study stands out in particular for its low synthesis contribution in figure 7.6:

- (H_7) **Neale *et al*.** (1998), entitled 'Fruit: Comparisons of attitudes, knowledge, and preferences of primary school children in England and Germany', aimed to examine differences in attitudes, knowledge and preferences according to gender, social class and culture. Data were collected via a questionnaire with mainly fixed response options and administered within an interview.

This study was judged to be one of the two lowest quality studies in the children and healthy eating review. Although the aims of this study were quite wide-ranging study findings were limited and only contributed to one of the 13 descriptive themes in the

healthy eating synthesis ('food preferences'). The study found that nearly all children regardless of sex or socio-economic background agreed that more fruit should be eaten and that favourite fruits included apples, strawberries, grapes and bananas (least favourite fruits were figs and dates). It is interesting to compare this study to Baxter *et al.* (2000) discussed above. The aims of these studies – exploring children's preferences for fruit or vegetables - were quite similar, yet the findings from Baxter *et al.* (2000) offer more depth and explanatory power. The findings of Neale *et al* (1998) and Baxter *et al.* (2000) both tell us which fruits and vegetables children prefer but it is only the findings of Baxter *et al.* (2000) which start to explain why children prefer some vegetables to others.

In contrast to Baxter *et al.* (2000), Neale *et al.* (1998) was badly reported; employed few techniques for increasing rigour in data collection and analysis; and used only one strategy - appropriate methods for collecting data from children - for ensuring that findings were grounded in the perspectives and experiences of children themselves (photographs and coloured stickers were used to help children express their views). Baxter *et al.* (2000) were also judged to have grounded their analysis methods (as well as their data collection methods) in the perspectives and experiences of children themselves. Some of the factors they analysed for an association with children's preferences were derived from children's own perceptions of fruit (e.g. sensory properties). However, Neale *et al.* (1998) only analysed children's preferences according to socio-demographics. This might explain the greater contribution that Baxter *et al.* (2000) made to the synthesis.

The second lowest quality study in the children and healthy eating synthesis also had a relatively low synthesis contribution score (although it was by no means the lowest):

- (H_8) **Tilston *et al***. (1991) entitled 'Dietary awareness of primary school children' aimed to examine children's understanding of healthy eating in relation to specific food items. Data were collected in group interviews on whether children thought they should eat more, less or the same of eleven different food items (e.g. bread, fruit, salt, sugar). Comments from children on food and healthy eating were also recorded.

Again this study was badly reported; employed few techniques for increasing rigour in data collection and analysis; and used only one strategy - appropriate methods for collecting data from children - for ensuring that findings were grounded in the perspectives and experiences of children themselves (games and picture cards were used to help children express their views). The main finding of this study - that the children's ratings on whether they should be eating more or less of particular food items did not always match nutritional messages - was fairly limited in terms of the review question. However, the reported comments from children were more illuminating (although it was unclear how these data had been analysed) and contributed to three of the thirteen descriptive themes (health consequences, food preferences and perceptions of health benefits). For example, the comment from one child in this study "I don't like them so they must be healthy" (p27) was used in combination with findings from other studies to suggest that children reject foods that are labelled as healthy or 'good for you'.

The two lowest quality studies in the physical activity review were also the two studies that contributed least to the synthesis results:

- (P_1) **Burrows *et al***. (1999) entitled 'Children's perceptions of exercise: are children mini adults?' aimed to examine whether children use the same "psychological constructs" (p 63) as adults in relation to taking part in exercise. Data were collected

using the draw and write technique (children were asked 'is there anything you would like to draw or write about exercise?')

- (P_2) **Davies and Jones** (1996) entitled 'Environmental constraints on health' aimed to examine children's perceptions of risk and patterns of decision-making on how they get around their local environment. Focus groups were used to collect data

The study by Burrows *et al.* (1999) met only five quality criteria. Its strengths lay in its attempts to use rigorous and appropriate methods of data collection to help children express their views (the draw and write technique). However, this study was very poorly reported and was judged to have used inappropriate data analysis methods for ensuring that findings were rooted in children's views because they used a pre-specified coding frame developed with adults to analyse data. Like some of the low quality studies with low contribution scores in the analysis of the reviews focused on young people, although the study aims were a close match to the review question, findings were limited in depth and detail. Findings were a list of barriers and motivating factors for exercise alongside the proportion of children mentioning each (e.g. x % of children reported that lack of time was a barrier to taking part in exercise)

Like Burrows *et al.* (1999), Davis and Jones (1996) only met five quality criteria. Reporting quality was poor; few techniques to increase rigour in data collection and analysis were reported; and due to lack of information, it was not possible to tell whether appropriate analysis methods were used to ensure that study findings were rooted in the views of children. However, in contrast to Burrows *et al.* (1999), the aims of this study were not a close match for the review question. Although this study did have some relevant findings on how the environment made it difficult for children to keep healthy and active (e.g. dangers of local environment such as

neglect of local play areas), the fact that this study had a much broader focus on risk perception and how children get around their local environment is a more likely explanation for why this study did not contribute as much as other studies to the synthesis.

Although the aims of Neale *et al*. (1998), Tilston *et al*. (1991) and Burrows *et al*. (1999) were a close match to the aims of the reviews they were included in their findings tended to be sketchy or limited in detail. It is this lack of detail in findings, rather than methodological quality by itself, that explains their low synthesis contribution. For Davis and Jones (1996), like the high quality/low contributing studies, it is a mis-match between the aims of the study and the focus of the review it was included in (their topic focus was peripheral to the topic focus of the review) that explains the low synthesis contribution.

<u>d) High quality studies with high synthesis contribution</u>

Three high quality studies stand out in figure 7.6 for their high synthesis contribution scores:

- (H_2) **Dixey *et al*.** (2001) entitled 'Children talking about healthy eating' aimed to examine children's understandings of healthy eating and influences on healthy eating in the context of obesity. Data were collected via focus groups.

- (H_6) **Mauthner *et al*.** (1993) entitled 'Children and food at primary school' aimed to explore children's views on, and experiences of, food and eating, with a particular focus on school meals, the social reasons for food choice and where experiences of food and eating fit within the context of children's whole lives. Interviews, group discussions and participant observation were used to collect data.

- (P_4) **Mulvihill *et al***. (2000b) entitled 'Physical activity at our time' aimed to explore a range of issues regarding children and physical activity including: barriers and motivations; preferences for different activities; and the role of parents and friends. Data were collected via focus groups

In contrast to the high quality/low contributing studies, the aims of these three studies were specified in more general and broad terms making them a close match to the aims of the reviews. Like the reviews, study questions were quite general (e.g. 'what are children's perspectives on, and experiences of, healthy eating?' as opposed to 'do children have the same attitudes to physical activity as adults?'). Reporting quality was very good in all three studies and each was judged to have made 'some' or a 'good' attempt to establish the reliability and validity of data collection methods. (This wasn't the case for data analysis methods however as reviewers did not record any attempt to increase the rigour of data analysis.) Another factor that makes these studies different to those high quality studies that had lower synthesis contribution scores (and to all of the other studies) is the fact that reviewers had judged their methods to be highly appropriate for studying children's perspectives and experiences.

The above is another reflection of how well the aims and methods of the three high quality/high contributing studies matched the focus of the review. These three studies shared many of the characteristics of the high quality/ high contributing studies in the analysis of the three reviews on young people's health above: a commitment to the value of studying the perspectives and experiences of children as a route to generating valuable knowledge; employing methods which are consistent with this commitment (e.g. ensuring that children feel comfortable to express themselves freely); and findings which went beyond a listing of the issues raised by children to the creation of explanations for the phenomenon or 'problem'

under study (e.g. why are there low levels of fruit and vegetable consumption amongst children?).

The study by Dixey *et al.* (2001) was undertaken alongside a trial of an obesity prevention intervention (the 'Apples Project'). Study authors argued that understanding children's views on healthy eating is essential for the development of interventions and that the reasons children may choose to eat healthily or not are likely to be different from adults. The rationale for the study by Mauthner *et al.* (1993) was that little is known about children's food choices and that most research on school meals has focused on parents and provider views. This study also aimed to develop suitable methods for research about children's perspectives and experiences. The study by Mulvihill *et al.* (2000b), included in the analysis of the three reviews on young people described earlier in this chapter, was also developed to inform an intervention (the 'Active for Life' campaign to promote physical activity). As noted earlier, with a focus on the views of children and young people, these authors expected their study to contribute new insights to the existing literature on factors influencing physical activity participation.

These study authors followed through on their commitment to studying the perspectives and experiences of children in their methods. In all three studies the analysis was driven by the children's perspectives rather than a priori frameworks (e.g. analysis was driven by questions such as what are children ideas about x and y?). Dixey *et al.* (2001) reported that: focus groups were used to collect data; techniques were employed to help all children express their views; and the focus group format was piloted with children of similar age to those in the full study. Mauthner *et al.* (1993) used a range of methods to collect data with children based on what the researchers had found to work in previous research and through observing and getting to know the children. For example, 'mini' focus groups were

used as the researchers found that these mimicked the small group classroom interactions in classrooms in which conversation between the children flowed. Mulvihill *et al.* (2000b) used paired interviews to collect data that incorporated the 'draw and write' technique to help children express their views. The authors report that the interview schedule was piloted prior to use and that children were allowed to exert an influence over the choice of topics that were talked about.

The findings of these studies were rich in detail and provided greater depth of insight into children's views on healthy eating and physical activity than those studies which, for example, listed the range of issues raised by children (e.g. Burrows *et al.*, 1999) or focused on a relatively narrow issue such as children's preferences for different fruits (e.g. Neale *et al.*, 1998). Although study authors did not report a great deal of detail on methods of analysis, the way that the data on children's perspectives and experiences was structured and used to explain the phenomenon under study suggested a thoughtful and rigorous process. For example, Mauthner *et al.* (1993) identified five factors to account for why children chose and ate the foods that they did at school: quality and type of food available; the quantity of food; children's personal preferences; cultural, economic and religious differences; and social settings and relationships. Like Mauthner *et al.* (1993), Dixey *et al.* (2001) identify factors that do and do not influence food choices (e.g. health consequences do not, taste preferences do) and concluded that children are active decision-makers in making food choices.

The 'richness' and 'depth' of the findings from the above studies were critical in the healthy eating review to moving beyond the 'descriptive' level of synthesis to the 'analytical' level. Dixey *et al.* (2001) and Mauthner *et al.* (1993) were the two studies to be given the most emphasis in the analytical part of the synthesis. These two studies were cited in support of more analytical themes than any of the other studies

in the part of the review report that descried the analytical themes (see Thomas *et al.*, 2003, pages 86 to 90)[5]. For example, as well as supporting analytical themes entitled 'future health consequences' (e.g. children prioritise taste over health) and 'fruit vegetables and confectionary' (e.g. fruit, vegetables and sweets mean different things to children), the findings from Mauthner *et al.* (1993) about the importance of social setting and social relationships for influencing food choice were also key to the emergence of 'children exercise choice' (e.g. eating sweets as a way of asserting independence from adults) and 'eating as a social occasion' (e.g. eating sweets to bond with friends).

e) Low quality studies with high synthesis contribution

One low quality study stands out in figure 7.6 for its relatively high contribution to the synthesis it was included in:

- (P_3) **Mason** (1995) entitled 'Young people and sport in England' aimed to explore children's views on sport and the personal and social influences on their participation. Interviews were used to collect data.

In the physical activity review the study by Mason (1995) made the biggest contribution to the synthesis. In the children and healthy eating review the study by Edwards and Hartwell (2002) stands out as a low quality study making a considerable contribution to the synthesis:

- (H_3) **Edwards and Hartwell** (2002) entitled 'Fruit and vegetables: attitudes and knowledge of primary school children' aimed to assess children's ideas about

[5] Mauthner *et al.* (1993) was cited within four of the six analytical themes and Dixey *et al.* (2001) was cited within three. Other studies were only cited once or twice (Baxter *et al.*, 2000; Edwards and Hartwell, 2002; Gibson *et al.*, 1998; Hart *et al.*, 2002; Neale *et al.*, 1998; Tilston *et al.*, 1991)

healthy eating as well as their knowledge about, and attitudes towards fruits and vegetables. Data were collected via interviews (to determine knowledge), discussion groups (to collect ideas about healthy eating) and self-completion questionnaires (to gather acceptability ratings for fruit and vegetables).

One might expect these studies to share some of the characteristics of the high quality studies with high synthesis contributions described above. The aims of these studies were a close match to the questions of the reviews they were included in. Mason (1995) aimed to explore children's views on participation in sport as a route to understanding the influences that affect their involvement and Edwards and Hartwell (2002) aimed to generate understanding about how children interpret the concept of 'healthy eating'. However, unlike the high quality studies with high synthesis contributions, it was not always clear whether study authors had followed through with this aim to the methods. In Mason (1995) data were collected via interviews but parents were present and were asked to help children answer questions. There is no detail on whether or how interviewers helped children to feel at ease and it is unclear whether the presence of parents had a positive or negative effect in this respect. The self-completion questionnaires used by Edwards and Hartwell (2002) were designed to be suitable for children and were tested for their appropriateness with children before the study started, but no details on the interviews or group discussions (which were also used to collected data) were presented.

Very little detail is presented on data analysis methods in either study. Edwards and Hartwell (2002) simply stated that 'qualitative' data were collated and sorted by age and response and Mason (1995) reported that interview transcripts were read and explored for main themes. Although Mason (1995) described the content of the interviews in detail and plenty of quotes from children and parents are presented,

the data appeared to have been under analysed. The reviewers commented that the findings of this study left the "reader with a very large amount of information to potentially take on board, almost requiring them to run their own data analysis". It is almost as if the study author took a decision to 'let the data speak for themselves'. Reviewers also commented that quotes from children and parents sometimes seemed at odds with the headings they were listed under or the summary points the author was using the quotes to illustrate.

Like the findings of the low quality studies with high synthesis contribution scores in the analysis of the reviews focused on young people, the findings of these studies did not display the conceptual depth and explanatory power of the high quality studies. The findings from Edwards and Hartwell (2002) were in the form of proportions of children correctly identifying fruits and vegetables, expressing a particular view about healthy eating, and stating a preference for different fruits and vegetables. As noted above, the findings of Mason (1995) were largely a description of the content of the interviews summarised under broad headings such as 'what children liked and disliked about sport'. Nevertheless, because the findings were relevant and detailed reviewers were able to extract a considerable number of findings from the studies. In the case of Mason (1995), this involved reviewers re-organising the mass of presented data into lists of barriers and facilitators of physical activity according to whether they were described by children or their parents.

f) Unique contributions

The only studies that contributed unique themes in the children and healthy eating review were two of the high quality studies H_2 (Dixey *et al.*, 2001) and H_6 (Mauthner *et al.*, 1993) (figure 7.7 and table 7.6) Without Dixey *et al.* (2001) or Mauthner *et al.*

(1993) just over half of the 13 descriptive themes would have been lost to the review: 'knowledge-behaviour gap' (e.g. children descried how their good intentions broke down in the face of temptation or easy access to sweets and biscuits); 'factors children describe as not influencing them' (e.g. advertising and friends); 'factors further constraining a limited choice at schools' (e.g. pressure to choose and eat food quickly); 'school dinners as a social occasion' (e.g. importance of sitting with friends); 'contradiction between promotion and provision of healthy foods' (e.g. healthy eating preached in the classroom, unhealthy foods provided in the canteen); and 'breaking rules and asserting independence' (e.g. choosing to eat sweets despite the rules). In contrast, very little would have been lost to the review if the low quality studies such as H_7 (Neale *et al.*, 1998) and H_8 (Tilston *et al.*, 1991) had been excluded from the review.

Table 7.6: Study contribution to the descriptive themes found in a synthesis of children's perspectives and experiences in a review about promoting healthy eating amongst children

	H_1	H_2	H_3	H_4	H_5	H_6	H_7	H_8
Awareness and understanding of healthy eating concepts		✓	✓		✓			
'Good' and 'bad' foods			✓		✓	✓		
Health consequences		✓	✓		✓	✓		✓
Food preferences	✓		✓	✓		✓	✓	✓
Perceptions of health benefits	✓	✓		✓				✓
Knowledge-behaviour gap		✓						
Roles and responsibilities		✓						
Factors children describe as not influencing them		✓						
Factors further constraining a limited choice						✓		
School dinners as a social occasion						✓		
Contradiction between promotion and provision of healthy foods		✓				✓		
Parental influence and food rules		✓			✓			
Breaking rules and asserting independence		✓				✓		

H_1 Baxter *et al.* (2000)
H_2 Dixey *et al.* (2001)
H_3 Edwards and Hartwell (2002)
H_4 Gibson *et al.* (1998)
H_5 Hart *et al.* (2002)
H_6 Mauthner *et al.* (1993)
H_7 Neale *et al.* (1998)
H_8 Tilston *et al.* (1991)

The picture is a bit different, however, for the children and physical activity review. In this review, the biggest loss to the review would occur if one of the low quality studies - P_3 (Mason, 1995) - had been excluded. Mason (1995) dominated the children and physical activity synthesis and twelve of the 39 barriers and facilitators would have been lost to the review had this study been excluded. These unique barriers and facilitators mainly focused on PE and sport in schools (e.g. frustration with unclear and/or complex rules for sports). This is something of an anomaly as there were no other examples amongst the reviews under study in this chapter in which a low quality study dominated the synthesis. There is no clear explanation for this anomaly but one possibility lies in the fact that a) there were only a very small number of studies in the children and physical activity synthesis and b) Mason (1995) was the only study to focus on physical activity at school. It could be that when syntheses include only one or two high quality studies and a small number of studies overall, a low quality study will be more likely to play a bigger role in a synthesis especially if it is the only study to report findings about one particular aspect of the phenomenon under study.

7.4 Discussion

The analyses reported in this chapter attempted to examine the relationship between the quality of 'qualitative' research and the findings of syntheses about intervention processes and people's perspectives and experiences. This relationship was difficult to study and turned out to be far from straightforward. In the first analysis, which focused on a synthesis of process evaluations, it was predicted that lower quality studies would produce different findings about the appropriateness of the intervention under study compared to higher quality studies, and that the findings of lower quality studies would be less reliable due to bias and error. The analysis revealed that lower quality studies were more likely to reach positive

conclusions about the appropriateness and effectiveness of peer-delivered health promotion interventions for young people. Higher quality studies were more likely to report mixed appraisals and refrain from drawing conclusions about effectiveness. This result suggests that study quality does indeed affect the results of process syntheses because lower quality process evaluations could mislead us about the appropriateness of interventions. However, the number of studies included in the analysis was small and further work would be needed to rule out the possibility that the result was due to chance or confounding factors. In addition, because the studies had been quality assessed largely on the basis of their reporting quality, it was not possible to be more specific about why low quality studies had different findings to high quality studies (e.g. failure to use techniques for increasing the rigour of analysis and/or inappropriate lines of questioning to collect data). Further work could test out some of these hypotheses.

In the second and third analyses, which focused on syntheses of children's and young people's perspectives on, and experiences of, various health topics, it was predicted that higher quality studies would contribute more than lower quality studies to the synthesis they were included in. This prediction was borne out to some extent. Although there were a number of exceptions, low quality studies did not contribute as much to the synthesis as the higher quality studies. Indeed, in all but one of the syntheses analysed, very little would have been lost to the review if low quality studies had been excluded. In contrast, syntheses would have ended up very bare if some of the high quality studies had not been included. These results suggest that lower quality studies of people's perspectives and experiences tend to offer a very limited or partial picture of the phenomenon under study. This result provides some much needed empirical evidence to inform the debate about whether or not studies should be quality assessed and then included or excluded on the basis of that assessment.

However, on closer inspection study quality – in particular reporting quality and methodological rigour - emerged as a secondary issue for making sense of why some studies contributed more than others to the syntheses they were included in. Regardless of reporting quality or methodological rigour, the aims and focus of the studies that made a high contribution to the synthesis they were included were a close match to the aims and focus of the reviews, and these studies also had very relevant and detailed findings. Nevertheless, quality did play a role when it was judged in terms of the appropriateness of study methods for answering the review questions about children's perspectives and experiences. High quality/high contributing studies were more likely to have used very appropriate methods for studying children's perspectives and experiences compared to low quality studies or high quality/low contributing studies. In other words, it is the relevance of study aims and the appropriateness of study methods to the question under review rather than reporting quality or methodological rigour that relates to how much a study will contribute to a synthesis on people's perspectives and experiences.

The finding that the appropriateness of methods for studying people's perspectives and experiences was a key factor for explaining synthesis contribution whilst reporting quality and methodological rigour were secondary factors provides some empirical support for the theoretical work of Popay *et al.* (1998). These authors argue that the primary marker of quality in 'qualitative' research – where 'qualitative' research is defined as studies that seek rich and deep data - should be whether or not the study has illuminated subjective meanings that shape action and behaviour by seeking understanding of a phenomenon from the point of view of a particular culture, society or group. Secondary markers are designed to help assess whether the primary marker has been met and cover flexibility of design, use of theoretical or purposeful sampling, adequate description, data quality, theoretical and conceptual adequacy, and typicality. The results of the analyses reported in this chapter

support the work of Popay *et al*. (1998) because the results suggest that, if studies do not use methods that privilege subjective meaning, they will be less useful in a synthesis even if they are well reported and executed.

Another major finding from the second and third analyses reported in this chapter was the importance of the 'form' (as opposed to content) of study findings for explaining synthesis contribution. Regardless of synthesis contribution, the only findings to display conceptual depth and explanatory power came from the higher quality studies. In contrast, the findings of lower quality studies (and some of the higher quality studies) came in the form of a) proportions of participants agreeing or disagreeing with a pre-constructed list of factors or b) a list of factors that participants had raised (sometimes accompanied by the proportion of the sample who mentioned a particular factor). These different 'forms' of findings had different roles to play in the two different types of syntheses conducted in the reviews analysed: the 'aggregative' style of synthesis that produced lists of barriers and facilitators across studies; and the 'interpretive' styles of synthesis which required the review team to move beyond concrete and descriptive themes to abstract and analytical ones. Whilst the survey-like findings of the lower quality studies (and some of the higher quality studies) were useful in both the 'aggregative' syntheses and the first descriptive stage of the 'interpretive' synthesis produced in the children and healthy eating review, it was the conceptual depth and explanatory power of the findings produced by some of the high quality studies that was crucial for the second analytical stage of the 'interpretive' synthesis. This finding suggests that for an interpretive synthesis it would be useful to appraise 'qualitative' and other types of studies of people's perspectives and experiences according to the conceptual depth and explanatory power of their findings. (The tool used to assess the quality

of studies in the syntheses on children's and young people's perspectives and experiences did not cover conceptual depth and explanatory power.)[6]

What the above suggests is that it is important to attend to the different forms of findings that 'qualitative' studies can produce when appraising and synthesising research. Sandelowski and Barroso (2003b) have made a similar point in relation to synthesis as a result of grappling with the methodological issues in their synthesis of 'qualitative' research about the experience of mothering amongst women diagnosed as HIV positive. They identified five categories of findings amongst their sample of studies: i) no findings; ii) topical survey; iii) thematic survey; iv) conceptual/ thematic description; and v) interpretative explanation. Like the analyses reported in this chapter, Sandelowski and Barroso (2003b) found that findings in the form of number ii) or iii) were only useful for a descriptive 'meta-summary' of findings across individual studies whereas findings in the form of iv) or v) were useful in both their descriptive 'meta-summary' and their interpretative 'meta-synthesis'. Whilst Sandelowski and Barroso (2003b) do not outline any implications of different forms of findings for quality assessment it seems highly plausible to suggest that quality markers may need to be tailored for different forms of findings. If study findings do not go beyond describing a list of issues raised by participants then it would seem inappropriate to assess such a study according to, for example, whether concepts or theory are well developed.

---

[6] A similar idea has been advanced by Popay and colleagues in relation to process evaluations. Popay *et al.* (2003, p 50) argue that process evaluations should be assessed according to the "explanatory quality of evidence on implementation" that they provide as well as methodological rigour (e.g. reporting quality, quality of design). They found that studies with greater power to explain the relationship between intervention implementation and outcomes included detailed descriptions of the intervention, its strengths and weaknesses, and its context; use of theory to build explanations; and the privileging of subjective meaning (Noyes *et al.*, 2005; Popay *et al.*, 2003). This is discussed further in chapter 8.

Diversity in the form of findings that 'qualitative' research can take is not well recognised in the literature on 'qualitative' research methods, either in the discussion of primary research or systematic reviews. For example, none of the tools proposed for assessing the quality of 'qualitative' research surveyed in chapter four make reference to the possibility that the findings of 'qualitative' research might come in different forms. The analyses reported in this chapter and the work of Sandelowski and Barroso (2003b) are therefore contributing a new perspective to the debate on assessing the quality of 'qualitative' research. In contrast to Sandelowski and Barroso (2003b), the results of this analysis also highlight that diversity in findings transcends the usual 'qualitative' and 'quantitative' distinctions. Findings displaying conceptual depth and explanatory power were not just the property of studies using methods traditionally associated with the label 'qualitative' studies (e.g. few variables specified in advance, open ended data collection methods), and findings in the form of, for example, proportions of the sample expressing particular views were not just the property of studies using methods traditionally associated with the label 'quantitative' studies (e.g. variables for analysis pre-specified in advance, data collected via fixed response options).

The findings of the analyses reported in this chapter need to be set against both the strengths and limitations of the study. As already noted, the first analysis on process evaluations was based on a small number of studies and the results should therefore be treated with caution. The finding that lower quality studies resulted in misleading findings about the appropriateness of interventions could have been due to chance. The second and third analyses were based upon a synthesis contribution score that represented the number of findings each study offered to the synthesis, with the expectation that higher quality studies would contribute more. A limitation of this score is that it reflects a rather crude assessment of synthesis contribution by putting 'quantity over quality'. This does not allow for the possibility that, whilst some

studies may not contribute much in terms of the number of findings, what they do contribute may be extremely useful or significant in the synthesis. Another way to study whether and how quality affects the results of syntheses that avoids these problems would be to sort studies according to high and low quality and conduct separate syntheses on each set of studies. The results of the syntheses of low quality studies could then be compared to the results of the synthesis of high quality studies.

Another limitation of the analyses reported in this chapter was that they were all retrospective. Whilst this offered the advantage of studying the relationship between study quality and synthesis results in a naturally occurring context, it also meant there was no chance to put safeguards in place to rule out error or confounding variables. For example, in the process used to conduct the review that formed the sources of data for this study, quality assessment was not done independently from the extraction of study findings. In other words reviewers were aware of the quality of studies when they selected relevant study findings for synthesis. It may be that reviewers were influenced by quality and were more cautious in their selection of findings from lower quality studies. This may account for the links between study quality and synthesis contribution observed in this study. Future studies could avoid this problem by using different reviewers to conduct quality assessment and synthesis.

Given these limitations a definitive answer to the question posed in the title of this chapter - whether the quality of 'qualitative' research affects the results syntheses of intervention processes and people's perspectives and experiences – remains elusive. However, with respect to syntheses of people's perspectives and experiences, the results of this study do suggest that excluding lower quality studies would not lead to a significant loss of findings. Furthermore, the attempt to answer

the question illuminated some important issues to consider in further work. Future work to assess the relationship between study quality and synthesis results requires consideration of the different dimensions of quality assessed; the different forms of findings that 'qualitative' studies offer; and the nature and purpose of the synthesis itself.

CHAPTER 8

Discussion and conclusion

8.1 Summary of thesis findings

This thesis has sought to advance knowledge about how to include and quality assess 'qualitative' studies in systematic reviews. Its starting point was a practical one: to meet the demand from evidence users that systematic reviews address issues of process, context, and need alongside effectiveness. Its findings, however, go beyond practicalities to contribute to fundamental debates about the purpose, methods and findings of research in the social sciences. The main body of the thesis consisted of a review of the conceptual and methodological literature on the topic of quality in 'qualitative' and 'quantitative' research and three new methodological studies. The first new study was a survey of existing tools for assessing the quality of 'qualitative' research, the second study analysed the development of a new tool for assessing the quality of 'qualitative' research, and the third new study examined the relationship between the quality of 'qualitative' research and the findings of systematic reviews.

The social science paradigm wars are a key influence on the existing literatures on how to assess the quality of 'qualitative' research and how to include 'qualitative' research in systematic reviews. In the paradigm wars 'qualitative' and 'quantitative' research are positioned as competing types of inquiry, with some 'qualitative' researchers arguing that 'quantitative' research is based on an inappropriate, outdated and 'positivist' model of science. This means that the major debates in the literature have been about whether the quality of 'qualitative' research can be assessed in the same way as 'quantitative' research and, within the realm of

systematic reviews, whether the 'quantitative' model of doing systematic reviews will fit 'qualitative' research. In the light of these debates, the findings of this thesis make three important new contributions to the literatures on how to include 'qualitative' research in systematic reviews and how to assess the quality of 'qualitative' research in systematic reviews and beyond.

The first contribution is a re-opening of the debate, closed down by the paradigm wars, about which criteria might be useful in the assessment of the quality of 'qualitative' research based on empirical data, not just epistemological or philosophical position. The new methodological studies in this thesis generated empirical data on the similarities, differences and usefulness of tools for assessing the quality of 'qualitative' research and on whether the quality of 'qualitative' research affects the results of systematic reviews that include them. The second contribution is a better understanding of how the 'quantitative' model of systematic reviews does and does not fit 'qualitative' research. One of the common themes to arise across the three new methodological studies was the importance of a) recognising the different forms of findings which 'qualitative' research can produce and b) engaging with the findings of 'qualitative' research in the quality assessment process as well as the methods. This is a challenge to the 'quantitative' systematic review template which requires reviewers to engage only with study methods until the final synthesis stage of the review. Given the inadequacies of the terms 'qualitative' and 'quantitative' for describing research activity, the third and final contribution of this thesis is a reframing of the debate about whether or not the 'quantitative' model of systematic reviews will fit 'qualitative' research to one about whether or not the model of systematic review developed to answer questions about the effects of interventions will fit questions about intervention processes and people's perspectives and experiences.

These three contributions represent the yield from the overall programme of work undertaken for this thesis. The rest of this section discusses the findings from the three new methodological studies individually. The common themes to emerge across the studies are also discussed. This discussion illuminates in more detail the three contributions from the overall programme of work as well what each study contributes individually.

i) Study one

The first methodological study was a survey and evaluation of existing tools for assessing the quality of 'qualitative' research. Tools were identified and then described, compared and evaluated for use in systematic reviews. Although a substantial number of tools were identified, the study findings were disappointing with respect to the study's practical aim: to provide a resource for reviewers who want to include 'qualitative' research in a systematic review. The study was not able to identify 'ready-made' tools that could be 'picked off the shelf' and used in a systematic review (or indeed for any other purpose that required 'qualitative' research to be assessed). Reviewers would encounter several problems with existing tools if they were to apply them to a report of a 'qualitative' study. Common practical problems included: a confusion in the tools between the study with the report that describes it; a failure to distinguish between items which asked reviewers to describe aspects of the study and those which required a quality judgement; a lack of guidance on how to make judgements such as 'adequate', 'sufficient' or 'key'; and a lack of provision for recording or making overall judgements on quality. These problems can be explained partly by the fact that few of the tools had been tried out in practice and then revised accordingly. Perhaps more fundamentally, tools tended to 'sit on the fence' with respect to quality. Few tools offered guidance on making an overall quality judgement and none directed reviewers to 'weigh up' the different

items in the tool to distinguish between minor flaws in the study and bigger problems which might be significant enough to question the trustworthiness of the findings of the study. In other words, reviewers would not be able to rely on existing tools to help them to decide whether to include or exclude a study from a synthesis on the basis of quality.

One reason for this was that there was no consensus across the tools on how the quality of 'qualitative' research should be assessed. Based on analysis of tool content, three types of tools were identified. *Methods-orientated tools* had a predominant focus on fieldwork, data collection, and analysis and were often very specific with regard to strategies to increase rigour and reporting. *Findings-orientated tools* included less detail on methods but had more extensive coverage of findings. Whereas methods-orientated tools focused on whether findings were supported by data, generalisable and useful, findings-orientated tools covered a greater range including: whether findings were clear and distinguishable, whether concepts or theory were well developed; whether diversity in meaning, perspective, and experience were captured; and whether findings resonated with readers and participants. *Methods- and findings-orientated tools* struck a balance between items about methods and findings, with some managing to retain a fairly detailed coverage of both domains. Across these three types of tools several dimensions of quality were represented. Some tool authors suggested that 'qualitative' research should be judged according to whether findings are 'valid' or 'accurate' accounts of the phenomenon under study. Other tool authors suggested that findings should be judged in terms of the 'vividness' of the descriptions produced about the phenomenon of interest or the 'analytical preciseness' of the explanatory theories. Yet others argued for findings to be judged in terms of their 'fertility' for generating new ideas or understandings or their 'utility' for addressing the practical or policy issues under investigation.

If tools for assessing the quality of 'qualitative' research are not suitable for practical application, this raises the question of why they were developed at all. Pawson (2006b, p113) characterised tools to assess the quality of 'qualitative' research as "a collection of tribal nostrums" about how to do good 'qualitative' research rather than as an aid to judging whether findings from 'qualitative' research can be relied upon. This description is certainly supported by this study which revealed over five hundred assessment items within the tools covering all possible aspects in the conduct and reporting of 'qualitative' research. The findings of this study also suggest that tools can be characterised as a manifesto on 'qualitative' research, designed to persuade readers that 'qualitative' research is as good as, or better than, quantitative research, and to convince us that 'qualitative' research is a special type of research, different in every way to 'quantitative' research. (Although, if one of the motivations behind the development of tools to assess the quality of 'qualitative' research has been to show how 'qualitative' research is different, it is ironic that a large majority of nearly all of the 545 assessment items proposed across the tools could be applied to *any* type of research.) In other words, tools to assess the quality of 'qualitative' research appear to be another site where the paradigm wars are played out. Playing out the paradigm wars could explain why tool authors have shown little interest in testing out their tools and why there is little agreement amongst researchers about which criteria to use. It has after all been over 20 years since Lincoln and Guba (1985) appealed to researchers to operationalise and try out the quality criteria they proposed. But today there is still a lack of empirical work.

The brief history of the development of tools for assessing the quality of trials described in chapter 2 suggests that, just like the situation for 'qualitative' research, early tools covered a huge range of issues covering, for example, trial organisation and ethics, as well trial design, implementation and analysis. Systematic review activity focused attention on quality related to trial design, implementation and

analysis because empirical evidence was accumulating to suggest that these things really mattered for producing the most reliable estimates of the effects of interventions[1]. Recent tools for assessing the quality of trials are therefore question driven and focused on examining the validity of the knowledge claims made in relation to that question. In contrast to current tools for assessing the quality of 'qualitative' research, these tools are not repositories of general wisdom providing comprehensive guidelines for good practice in designing, implementing and analysing trials. They focus on one specific issue: whether the effect size produced by the trial is a reliable estimate of the effect of an intervention. Accordingly, individual items in the tool focus on assessing to what extent the design, implementation and analysis of the trial has minimised the introduction of bias and error into the effect size estimate (e.g. through non-random allocation of participants to intervention and control groups).

Another way of describing the focus of quality assessment tools for trials is that they help the reviewer to identify the "fatal flaws" in a trial (Dixon-Woods *et al.*, 2004a, p9). ('Fatal' here refers to those flaws which cast doubt on the findings of the trial as the best approximation of the true effect of an intervention). The growing interest in doing systematic reviews which ask questions requiring the inclusion of 'qualitative' research provides an opportunity to examine the issue of quality in 'qualitative' research in a new way, one which is question rather than method or epistemologically driven, and one which allows us to step outside of the paradigm wars. A promising area for the further development of tools to assess the quality of 'qualitative' research is to try to identify what the 'fatal flaws' might be.

---

[1] This is not to say that ethical and other issues such as relevance are not important. All research should strive to be ethical and relevant as well as scientific.

b) Study two

The second methodological study in this thesis analysed the development of a new tool that could be applied to assess the quality of 'qualitative' research. The final version of this tool was significantly different to the tools identified in study one because it created a link between the review question under study and the individual quality assessment items. The tool was question rather than method or epistemologically driven. This tool required an overall assessment of study quality which directly addressed whether the study findings could be relied upon to answer the review question, and individual tool items about how well the study was carried out were linked to this overall assessment. The analysis revealed several factors which were key influences on the development of this new approach to judging the quality of 'qualitative' research. Interestingly, a desire to follow scientific principles was only part of the picture, the backgrounds, interests and principles of the individuals involved in the development, the way these individuals worked together as a team, and the wider social context in which the development took place also emerged as significant factors. These factors are usually hidden or implicit in accounts of research. Indeed study one found that detailed accounts of how tools were developed were rare amongst existing tools to assess the quality of 'qualitative' research. This is quite surprising given the emphasis on reflexivity amongst 'qualitative' researchers.

The team working on the development was a multi-disciplinary one covering biological sciences, nursing, geography, psychology, sociology, social policy and education. Working in a team contrasts with the 'lone researcher' model and is thought to offer several advantages because it can facilitate the critical debate essential for rigorous and innovative research (Fenton *et al.*, 2001; Wray, 2002). Working in a multi-disciplinary team is considered to be especially useful in some

areas of research as it can facilitate lateral thinking between different disciplines and can protect against discipline 'blind spots' (Fuqua *et al.,* 2004; Hulme and Toye, 2006). This study found that the multi-disciplinary aspect of the team did play a role in the methodological development under analysis. Discipline specific assumptions were exposed and challenged and the team often had to go 'back to basics' to define terms which were taken for granted within individual disciplines. For example, the term 'qualitative' research was problematised as a useful way of describing the types of research relevant to studying people's perspectives and experiences, and the shorthand term 'views studies' was adopted by the team instead.

However, study two also found that it was important not to downplay the importance of the things that team members had in common for the methodological development. For example, team members all shared a commitment to the scientific method and trying out the tool in several reviews and then revising it was seen as an essential part of development. Furthermore, the team shared a commitment to both randomised controlled trials and to 'qualitative' research. This is a very unusual position for social researchers to adopt as these types of research have been constructed as opposing paradigms in research. The team operated on a principle of 'fit for purpose' whereby choice of research method is driven by the research question in hand rather than ideological concerns. In other words the team actively resisted and challenged the paradigm wars. This meant that none of the team came with a 'manifesto' on 'qualitative' research to impose at the expense of other types of research. These shared principles and understandings were also important for getting on with the job in hand.

The particular funding and policy climate in which the methodological development took place emerged as a significant factor at play within the wider social context of the work. Unlike the majority of tools identified in study one this tool was developed

in order to complete substantive systematic reviews to address policy issues. Working with real questions provided the team with a concrete framework in which to explore quality issues without getting too bogged down with philosophical debates or with trying to produce a tool to cover every aspect of the research process. The review questions provided an 'anchor point' for the research team to return to for setting priorities and to reflect upon work completed. It is important to note that doing research for policy did not mean that the team were not interested in broader debates about quality and the nature and purpose of 'qualitative' research in the social sciences. The interaction between the interests of the research team and the funding and policy climate is illuminating here. Including 'qualitative' research in systematic reviews had caught the intellectual imagination of the team. As well as being policy driven, the team were also driven by a desire to contribute to the wider debates about the nature and purpose of 'qualitative' and other types of research in the social sciences.

c) Study three

Influenced by the work of those who have studied whether methodological flaws present in trials impact on conclusions about the effects of healthcare (e.g. Chalmers *et al.*, 1983; Juni *et al.*, 2001b), the third and final methodological study in this thesis attempted to generate empirical evidence on the relationship between the quality of 'qualitative' studies and the findings of reviews. In other words, this study attempted to address the question of whether and how study quality affects the results of systematic reviews which include 'qualitative' research. The data for the study had been generated in a series of systematic reviews which conducted syntheses of 'qualitative' and other types of studies of intervention processes and people's perspectives and experiences. These studies had all been quality assessed using versions of the new tool which was the subject of analysis in study

two. The analysis in study three attempted to examine the relationship between each study's quality score and its findings in relation to the synthesis it was included in. Like methodological studies about trial quality, the analysis started out from the assumption that the findings of lower quality studies are subject to bias and error. The relationship between quality and findings proved to be a very difficult issue to study and, perhaps unsurprisingly, one the major findings of study three was that the relationship was not a straightforward one. However, several interesting patterns emerged suggesting that quality does indeed matter for systematic reviews which include 'qualitative' and other types of studies on intervention processes and people's perspectives and experiences. Different insights were gained from the first part of the analysis within the review which looked at intervention processes compared to the second part of the analysis within the reviews which produced syntheses of people's perspectives and experiences.

The first part of the analysis on the relationship between study quality and review findings for process evaluations focused on two issues: the acceptability of peer-delivered health promotion for young people; and whether study authors drew conclusions which were not justified by the study design employed. The analysis revealed that lower quality process evaluations were a) more likely to reach positive conclusions about the acceptability of peer-delivered health promotion and b) more likely to reach unwarranted conclusions that peer-delivered health promotion was effective. On the first issue it was not clear why poor quality studies were more likely to report only positive appraisals of peer-delivered health promotion. Because the quality of process evaluations was largely assessed in terms of reporting quality it was impossible to tell whether, for example, the authors had failed to use techniques for increasing the rigour of analysis such as searching for negative cases or whether they had used selective lines of questioning to collect data that made it difficult for

young people to express any negative views[2]. On the second issue, it was also not clear why unwarranted conclusions were drawn about the impact of interventions. One possible reason could be a lack of consensus about which methods can produce reliable findings about the effects of interventions (Oakley, 1998; Davies and MacDonald, 1998). The politics of funding could be another explanation. Funders may have wanted to know whether the intervention worked or not but did not provide the means to carry out the most appropriate evaluation to answer this question.

Like reviews of the effects of interventions, the review question for the synthesis of process evaluations required a single answer in the form of a 'yes' or a 'no' (Is peer-delivered health promotion appropriate for young people?). In contrast, the review questions for the series of reviews which included studies of people's perspectives and experiences required answers in the form of a list (e.g. What are the barriers to, and facilitators of, healthy eating amongst young people?). Accordingly, the second part of the analysis for study three, which examined the relationship between study quality and review findings for studies of people's perspectives and experiences, focused on what each study contributed to the synthesis it was included in. Here, the assumption was that lower quality studies would not contribute as much to the synthesis as higher quality studies. In the first four reviews in this series, the syntheses products were lists of barriers and facilitators, and the major finding here was that lower quality studies tended not to contribute as many barriers and facilitators to the syntheses as compared to the high quality studies. In the last two reviews of this series, the synthesis products were lists of themes, and again the major finding was that low quality studies did not contribute as much as higher

---

[2] Only one of the criteria required a judgement on how well the study was carried out and this concerned whether or not two researchers had carried out the data analysis. This criteria did not appear to be related to whether or not process evaluations had found only positive appraisals of peer-delivered health promotion.

quality studies to the generation of the themes. What is suggested by these findings is that compared to high quality studies, low quality studies tend to offer a partial or limited picture of people's perspectives and experiences. In reviews focused on understanding people's perspectives and experiences there may be little to gain from including lower quality studies.

More detailed analysis, however, revealed that study quality was in fact a secondary factor for explaining synthesis contribution. The primary factor was how close a match the study aims and focus were to the aims and focus of the review question. Regardless of reporting quality or methodological rigour, when the aims and focus of a study were a close match the review aims and focus, the study made a high contribution to the synthesis it was included in. Quality *did* play a role in explaining synthesis contribution when it was judged in terms of the appropriateness of study methods for answering review questions about people's perspectives and experiences. Those studies that were judged to have used highly appropriate methods for examining people's perspectives and experiences made the biggest contribution to the syntheses that they were included in.

The 'form' of study findings (as opposed to content) also played a role in explaining synthesis contribution. Findings came in one of three forms, the first were summaries or lists of the issues raised by participants (sometimes accompanied by proportions of participants raising each issue); the second were lists of issues put forward by researchers, ranked according to the number of participants agreeing that a particular issue was important or significant for them; and the third were in-depth descriptions of the phenomenon under study, often structured according to themes or concepts. It was only high quality studies that produced findings of the latter type. These different forms of findings had different functions within the different types of syntheses – aggregative and interpretive - conducted in the

reviews of children's and young people's perspectives and experiences. Studies with survey-like findings were useful in the aggregative syntheses (which produced lists of barriers and facilitators) and the descriptive stage of the interpretive synthesis, but it was only studies whose findings displayed conceptual depth and power that were useful for the analytical stage of the interpretive synthesis (which produced a set of descriptive and analytical themes).

d) Common themes across studies

A number of common themes emerge in trying to explain and make sense of the major findings across studies one to three and these have an important bearing on the original aim of this thesis – to advance knowledge about how to include and quality assess 'qualitative' research in systematic reviews.

The first is that the presentation of 'qualitative' research as a unified and coherent enterprise distinct from 'quantitative' research is a problematic one. An explanation of the finding from study one that existing tools for assessing the 'quality' of qualitative research are not very useful is that these tools actually represent steps in conducting the social science paradigm wars rather than aids to distinguish between high and low quality studies. Study two found that a key factor in the development of a more useful tool to assess the quality of 'qualitative' research was an ability to step outside of the paradigm wars and get down to the business of trying to link study quality to the trustworthiness of findings for answering research questions. One significant reason for why the relationship between study quality and the findings of systematic reviews proved so hard to examine in study three was because, in contrast to the homogenous presentation of 'qualitative' research in existing tools, the findings of 'qualitative' studies came in diverse forms. The importance of different forms of findings was not raised by existing tools and the new tool

described in study two had not taken these different forms of findings into account either.

The proposition that 'qualitative' and 'quantitative' research are problematic concepts is not a particularly new argument. As described in chapter one, other researchers have problematised definitions of 'qualitative' research when they are given in opposition to 'quantitative' research. However, as the methodological literature on systematic reviews extends beyond the boundaries of trials, there is evidence that the social science paradigm wars are being re-created in the literature on research synthesis. The phrases 'qualitative systematic reviews' and 'quantitative systematic reviews', for example, have gained broad and uncritical acceptance in some circles (e.g. Barbour and Barbour, 2003; Booth, 2001; Dixon-Woods *et al.*, 2006; Jones, 2004). The findings of this thesis represent a challenge to this acceptance and suggest that the debate associated with such acceptance - on whether the model of the 'quantitative' systematic review fits 'qualitative' research - should be reframed as whether the model of systematic review developed to aggregate the findings of trials to answer questions about effectiveness fits systematic reviews which include other types of research to answer questions beyond effectiveness such as those about intervention processes and people's perspectives and experiences.

The second common theme amongst the findings of studies one to three – that there is no straightforward or easy way to distinguish between 'qualitative' and 'quantitative' studies - adds further weight to the need to reframe the above debate. For example, the variations in the form of study findings identified in study three did not mirror the traditional 'qualitative' and 'quantitative' divide. Amongst studies of people's perspectives and experiences, findings which came in the form of proportions of participants expressing a particular view were not just a feature of

studies using methods associated with 'quantitative' research. To analyse fully the relationship between the quality of 'qualitative' research and the findings of systematic reviews in study three required unpacking the precise nature of the 'qualitative' or 'quantitative' methods used in each study. The fact that the majority of items in the existing tools for assessing the quality of 'qualitative' research examined in study one could be applied to *any type* of research also illustrates the difficulties in drawing a clear line between 'qualitative' and 'quantitative' research. The new tool had side-stepped these difficulties because quality assessment was linked to the review question. The team who developed the tool never intended the tool to be applied to 'qualitative' studies only. The final version of the tool was intended for any type of study that examined people's perspectives and experiences.

The third and final common theme which can help explain the findings of studies one to three is the lack of an empirical base to underpin methods for assessing the quality of 'qualitative' research. There are two areas in which empirical data are lacking. The first is on how tools for assessing the quality of 'qualitative' research work in practice. Existing tools for assessing the quality of 'qualitative' research were rarely tested and then revised. The second area in which empirical data are lacking is on which of the hundreds of quality assessment items which have been suggested over the years really matter in 'qualitative' research i.e. which ones represent the 'fatal flaws' that would render the findings of a 'qualitative' study as untrustworthy? This means that the choice of items used in quality assessment tools has so far been largely subjective (Attree and Milton, 2006). Moreover, it appears that the choice of items in existing tools have been driven by epistemological concerns rather than by research questions or by theoretical propositions about how study quality might affect the trustworthiness of findings in relation to particular questions. Although the team who developed the new quality assessment tool in

study two were driven by research questions, the findings of study three suggest that the next stage of development for quality assessment tools is to generate a set of hypotheses about what the 'fatal flaws' might be in studies of intervention processes and people's perspectives and experiences. One important reason for why the relationship between study quality and the findings of systematic reviews was so hard to study (and turned out to be far from straightforward) was the possibility that study quality had not been assessed in quite the right way. In fact study quality was assessed largely on the basis of reporting quality which may not reflect how well a study has been carried out. (Although the fact there was some relationship suggests that reporting quality might well be a useful proxy or indicator for how well the study was carried out and hence how well bias and error in the study findings were minimised).

8.2 How this thesis relates to and extends the work of others

In the literature reviewed in chapter 3 on quality in 'qualitative' research, assessing the extent to which bias and error had been minimised in studies was central to early ideas on how the quality of 'qualitative' research might be assessed. With the influence of post-modernism in the social sciences, however, these ideas became very unfashionable. The concepts of 'bias', 'validity', and 'reliability' were rejected by those adopting a relativist position. They argued that such concepts were inappropriate under the assumption of multiple truths and realities and that method should not be the sole route to assuring quality. (Some of those writers taking this perspective argued for the quality of 'qualitative' research to be judged on its ethical or practical merit.) Subsequent literature on the topic has tended to get stuck within this debate, as authors position their own or other people's quality criteria as 'conventional' or 'alternative' (Murphy *et al.*, 1998); 'positivist' or 'post-positivist' (Devers, 1999); 'naïve realist', 'subtle realist', or 'interpretivist' (Angen, 2000); and

'realist'; 'contextualist'; or 'radical constructionist' (Madill *et al.*, 2000). Somewhere along the way the challenge to test out quality criteria seems to have got lost and the debate about which quality criteria might be useful based on empirical data closed down before it had chance to begin. This thesis has therefore made an important contribution because it has re-opened this debate in its attempt to generate empirical data about how useful quality assessment tools are and whether the quality of 'qualitative' research matters in relation to the questions and findings of systematic reviews.

The diversity of findings in 'qualitative' research is another important finding from this thesis which extends the work of others. Although others have highlighted the diversity in 'qualitative' research in terms of methods and/or epistemological position, few have written about diversity in 'qualitative' research in terms of forms of findings. Interestingly, those who have engaged with different forms of findings have been those who are interested in using the findings of 'qualitative' research to inform policy and practice (Kearney, 2001; Sandelowski and Barroso, 2003b). Based on 'qualitative' research in nursing, Kearney (2001) created a typology of findings with five categories on a continuum according to the level of discovery and complexity of the findings. These categories ranged from 'findings restricted by a priori frameworks' through to 'dense explanatory description'. Sandelowski and Barroso (2003b) developed a similar typology on the basis of 'qualitative' research on women's experience of living with HIV. The five categories in this typology also rested on a continuum based on the extent to which data had been analysed. Categories ranged from 'no findings' (in which no analysis has been done and quotes from participants are simply presented 'in order to let the data speak for themselves') through to 'interpretive explanation'.

The forms of findings identified by this thesis within a different topic area (health promotion and public health) show remarkable overlap with both these typologies. The first type of finding identified in this thesis - lists of issues put forward by researchers, ranked according to the number of participants agreeing that a particular issue was important or significant for them - is similar to the category 'findings restricted by a priori frameworks' put forward by Kearney (2001). The second type of findings identified in this thesis - summaries or lists of the issues raised by participants - is similar to the 'descriptive categories' of Kearney (2001) and the 'topical survey' of Sandelowski and Barroso (2003b). The third type of finding in this thesis - in-depth descriptions of the phenomenon under study, usually structured according to themes or concepts – covers the 'shared pathway or meaning', 'depiction of experiential variation', and 'dense explanatory description' categories of findings from Kearney (2001), and the 'thematic survey', 'conceptual/ thematic description', and 'interpretive explanation' categories from Sandelowski and Barroso (2003b).

This thesis therefore supports and extends the work of Kearney (2001) and Sandelowski and Barroso (2003b). It is important to note however, that the typology produced by this thesis differs from the other two because it can be applied to both 'qualitative' and 'quantitative' studies of people's perspectives and experiences. Another difference is that the typology of this thesis makes no assumption about the quality of research within the different categories. In the typologies of Kearney and Sandelowski and Barroso, there is an assumption that categories one and two are of poorer quality or a lesser value than other categories (Kearney, 2001) or that categories one and two are not 'proper' 'qualitative' research (Sandelowski and Barroso, 2003b). The findings of my thesis do suggest that different forms of findings may need to be assessed against different quality markers. For example, if study findings do not go beyond describing a list of issues raised by participants

then it would seem inappropriate to assess such a study according to, for example whether concepts and theories are well developed.

Since 2001 there has been a rapid expansion of the literature on systematic reviews and 'qualitative' research, and some of this has focused specifically on quality. As described in chapter 3, this literature is important because it raises a number of debates about assessing the quality of 'qualitative' research in systematic reviews, although only a small proportion of this literature is based on empirical study. One such debate is whether quality should be assessed prior to the synthesis stage of a systematic review and whether these judgements should lead to the exclusion of studies on the basis of quality. Some argue that quality should be assessed before synthesis in order to ensure that only the best evidence is used (Attree and Milton, 2006; Campbell *et al.*, 2003). Others, however, question whether quality needs to be assessed prior to synthesis and suggest that the quality of studies will only emerge in the synthesis (Noblit and Hare, 1988; Pawson, 2006b).

This thesis lends support to the former position in this debate because it found that there is a relationship between the quality of studies and the findings of systematic reviews which include 'qualitative' research and that lower quality studies tend not to contribute as much as higher quality studies. However, because there is not yet a full understanding of this relationship, or of the fatal flaws in 'qualitative' research, it would be unwise to rush into a blanket exclusion of studies on the basis of quality. Indeed, studies of varying quality need to be synthesised in order to study whether quality makes a difference. And, in order to study whether quality makes a difference, a greater understanding of synthesis methods for the findings of 'qualitative' research is needed. In this respect it is important to remember that in the history of the development of systematic reviews of trials, methods for quality

assessment ante-dated methods for synthesis, and it was only after many syntheses had been done that quality assessment methods could be empirically tested.

As already noted in this chapter, one of the underlying debates in the literature on how to include 'qualitative' research in systematic reviews is whether the 'quantitative' model of systematic reviews will fit 'qualitative' research. In terms of quality assessment, a 'quantitative' model suggests that quality should be assessed in terms of how well the study was carried out without reference to study findings. Ideally, in systematic reviews of trials, reviewers are encouraged to assess methodological quality without reference to study findings to prevent more favourable assessments of those trials with 'desirable' findings (Alderson *et al.*, 2005; Cooper and Hedges, 1994). The findings of this thesis suggest that this ideal from the 'quantitative' model does not fit easily with 'qualitative' research. Study one found that the most useful tools for assessing the quality of 'qualitative' research were those that included items which required engagement with study findings as well as study methods, and study three found that the form of findings was an important part of the relationship between study quality and synthesis contribution. From a systematic review perspective the weight given in tools to assessing quality via study findings is an unexpected challenge. Like the 'conventional' and 'alternative' views on quality in 'qualitative' research described in chapter three, in the 'ideal type' systematic review there is an expectation that quality will be assessed according to the quality of study methods rather than findings.

The above suggestion that reviewers may need to engage with the findings of 'qualitative' research as well as methods for quality assessment is consistent with those of other recent methods work in this area which has begun to examine how quality assessment works from the 'inside' of specific reviews. Campbell and colleagues in the UK, who undertook a synthesis of 'qualitative' research on lay

experiences of diabetes and diabetes care using meta-ethnography, found that the most useful quality appraisal items in their tool were those that required reviewers to engage with the findings of studies: a) that the concepts and interpretations should be grounded in the data gathered; and b) that the concepts and interpretations posed were cogent and original (Campbell *et al.*, 2003). Similarly, Sandelowski and colleagues in the US, who undertook a synthesis of 'qualitative' research on women's experiences of living with HIV found the most useful appraisal items to be those in the 'findings' category which included items such as "concepts or ideas are well-developed and linked to each other" and "the results offer new information about, insight into, or formulation of the target phenomenon". Recommending engagement with the findings of 'qualitative' research for quality assessment is taking on board some of the messages from the 'radical' position on quality in 'qualitative' research (see chapter 3), which suggests that we should abandon judgements of quality based on methods. However, I am not suggesting that the 'radical' position should be taken up in its entirety. Posing 'cogent' and 'original' concepts and interpretations may be entirely independent from the employment of rigorous methods to minimise bias and error.

The findings of this thesis about the importance of considering study findings in quality assessment, as well as methods, resonate with the observations of Popay and colleagues in the UK who have studied how to include process evaluations in systematic reviews (Arai *et al.*, 2005; Noyes *et al.*, 2005; Popay *et al.*, 2003; Popay *et al.*, 2006). Popay *et al.* (2003, p50) argue that 'qualitative' process evaluations should be assessed according to the "explanatory quality of evidence on implementation" that they provide, which require a reviewer to engage with the findings of studies, as well as methodological rigour (e.g. reporting quality, quality of design). The findings of this thesis support the 'substantive' approach to quality assessment suggested by Eakin and Mykhalovskiy (2003). In a 'substantive'

approach to quality assessment, quality judgements "would emerge from a deeper engagement in and understanding of the interpretations and propositions being put forward and assessment of how they are (or are not) produced and rendered convincing by the research practices used" (Eakin and Mykhalovskiy, 2003, p192). Study appraisal would therefore not be restricted to reporting quality, methodological rigour and appropriateness of methods for the research question, but would extend to a consideration of the conclusions of the research and whether these conclusions were warranted given the methods used.

8.3 Strengths and limitations

I have already discussed the particular strengths and limitations of each of the three new methodological studies in chapters five, six and seven. In this section, I discuss those strengths and limitations which cut across more than one part of the thesis.

a) Strengths

A key strength of the methods used in this thesis is the application of methods from both the 'quantitative' tradition and the 'qualitative' tradition. The use of 'mixed methods' has been advocated as a way to gain a more complete picture of the phenomenon under study than can be gained by either 'qualitative' or 'quantitative' methods in isolation (Barbour, 1999; Cresswell, 1995; Ragin, 1987; Tashakkori and Teddlie, 1998). Myself and colleagues have argued elsewhere that one way in which mixing methods provides a more complete picture of the phenomenon under study is through the use of different methods to answer different questions about that phenomenon (Harden and Thomas, 2005). In this thesis I used mixed methods to answer several questions about different aspects of the problem under study: how to include and quality assess 'qualitative' research. In study one, for example, I used

frequencies to find out how many different criteria had been proposed to assess the quality of 'qualitative' research in my sample of tools and then compared, contrasted and grouped these different criteria to identify themes that characterised the content of these criteria. As well as this "variable-orientated" approach to looking across tools, I also used 'qualitative' and 'quantitative' methods to compare and contrast individual tools in a "case-orientated" approach to analysis (Sandelowski and Barroso, 2002, p37). This involved establishing the weight given by each tool to the different study domains the quality criteria covered (e.g. background theory and research questions; findings) by calculating the proportion of the total number of tool items in each domain. I was then able to compare and contrast tools with different weightings to build up a picture of tool features that were associated with different weightings.

Another strength of this thesis lies in my use of several strategies to enhance rigour and minimise the introduction of bias and error into my findings. When possible I chose research designs that were the most appropriate for the questions under study, used standardised data collection tools (when the data to be collected could be specified in advance), made use of software packages to facilitate data storage and analysis, and carried out checks on my data analysis. For example, in study one I used a cross-sectional survey design to generate a description of available tools for assessing the quality of 'qualitative' research and used EPPI-Reviewer and NVivo to store and analyse data. Strategies for increasing rigour in study two included gathering detailed data on the development of the new tool for assessing the quality of 'qualitative' research, getting immersed in that data through reading and re-reading relevant documents and writing notes, and getting other members of the team who developed the tool with me to check my reconstruction of events. Strategies for increasing rigour in study three included the use of statistical analysis to test visual representations of the relationship between study quality and synthesis

results, and attention to negative or extreme cases to revise and refine my analysis to explain why 'qualitative' studies varied in the contribution they made to the syntheses they were included in. There were, however, some other strategies that I did not use for increasing for rigour that would have been highly appropriate. The retrospective nature of some of my analyses was also a problem. These issues are discussed next as limitations of my thesis.

b) Limitations

Although I was able to use several techniques for enhancing rigour, one technique I was not always able to employ was to have another researcher to conduct analyses with me or to challenge my emerging analyses. Use of two researchers to conduct analyses and/or the use of a critical peer to check emerging analyses have been recommended as way to increase the rigour of research (e.g. Beck, 1993; Elder and Miller, 1995; Malterud, 2001; Mays and Pope, 1995; Miles and Huberman, 1994). As already noted in chapter five, I would have felt more confident in my characterisation of tools to assess the quality of 'qualitative' research had another researcher been able to look in-depth at all or part of my data. A particular problem in my thesis was the fact that I came to the subject matter with a particular perspective on research. Unlike some of those who consider themselves a 'qualitative' or a 'quantitative' researcher, I see different methods as complementary rather than competing, but see some methods as better for answering particular questions than others. Although I have skills in both 'qualitative' and 'quantitative' methods I do not consider myself to be a 'quantitative' researcher or a 'qualitative' researcher. Whist such a perspective has its advantages, I do not necessarily have specialised knowledge within either tradition (e.g. the assumptions and techniques of multiple regression or the assumptions and techniques of grounded theory).

One possible consequence of this is that I may have overlooked factors which those with different perspectives would see as significant. For example, in study one I considered that nearly all of the items proposed by tools to assess the quality of 'qualitative' research could in fact be applied to *any* type of research. It would be interesting to see if those who identify as 'qualitative' or 'quantitative' researchers would agree with this. The perspective of another researcher may also have been useful to counter any bias, error or oversights introduced by an 'insider perspective' to the analysis of the development of a new tool to assess the quality of 'qualitative' research. As discussed in chapter six, because I was part of the team which developed the new tool, I may have wanted to cast that team in a positive light perhaps missing relevant and illuminating points of tension within the team.

The second and third methodological studies in this thesis were both retrospective in design. The development of the new tool to assess the quality of 'qualitative' research had already taken place so I had to use a retrospective design to analyse the factors influencing the development in study two. It is hard to imagine, however, how this study could have been done prospectively, unless one specific factor was selected (e.g. multi-disciplinary team) and then tested for its influence on the methodological development of teams with and without the factor. My analyses were also retrospective in study three because the reviews that I used as sources of data had already been completed. The retrospective nature of this study meant that there was no chance to put safeguards in place to rule out error or confounding variables in any relationship between study quality and systematic review results. However, as Dixon-Woods *et al.* (2004a) note, because trying to answer the question of how to include 'qualitative' research in systematic reviews is breaking new methodological ground, even basic reflections to share experiences on attempts to try to include 'qualitative' research will help to develop methods. Despite being

retrospective, my analyses went one step further than reflection by interrogating existing data sets in a systematic and rigorous way.

My intention in this thesis was to generate findings about assessing the quality of 'qualitative' research that would be relevant to any discipline or field of public policy. However, much of the material under analysis in study two and study three was from the field of health promotion and public health (HP&PH). In study three the data for analysis came from the studies included in reviews in HP&PH, and in study two the new tool to assess the quality of 'qualitative' research that I analysed was developed to apply primarily to studies in HP&PH. Even though the intention in study one was to survey tools for assessing the quality of 'qualitative' research from across the social and health sciences, the majority of tools in my sample were health-related. It may be, for example, the quality of 'qualitative' research in HP&PH is poorer than in other areas. (One part of the analysis in study three revealed that 'qualitative' research in HP&PH often failed to meet even very basic standards of reporting.)I therefore cannot be certain that the findings of this thesis would be the same if I had analysed, for example, systematic reviews conducted in different topic areas such as education, social welfare, or crime and justice. The inclusion of data generated in other areas of public policy would have greatly strengthened my claims for generalisability. Nonetheless, work examining the quality of 'qualitative' research in other areas such as education and social welfare do show similar problems to those identified in 'qualitative' research in HP&PH (Long and Godfrey, 2004; Oakley, 2003; O'Conner *et al.*, 2001). (In fact, the paper by Oakley, 2003 suggests that quality may be even worse in education). It does not, therefore, seem unreasonable to suggest that the findings of this thesis about assessing the quality of 'qualitative' research are likely to be relevant across disciplines and public policy fields.

8.4 Implications

The implications from this thesis fall into three areas: a) for systematic reviews which include 'qualitative' research; b) for future methodological work in this area; and c) for primary research.

a) Implications for systematic reviews which include 'qualitative' research

To date there has been little guidance for including 'qualitative' research in systematic reviews from either the key methodological textbooks in this area or the major national and international organisations which produce systematic reviews. The key textbooks on systematic reviews in healthcare focus only on statistical meta-analysis of trials and observational studies and do not even mention 'qualitative' research (e.g. Egger and Davey-Smith, 2001). Similarly, the major textbook in the social sciences, the *'Handbook of Research Synthesis'* (Cooper and Hedges, 1994), fails to mention 'qualitative' research. The same situation occurs within the handbook of the international Cochrane Collaboration (Higgins and Green, 2006) and the documentation offered by the international Campbell Collaboration, although these organisations have both set up specific methods groups to develop guidance. The Centre for Reviews and Dissemination in the UK has included sections on 'qualitative' research in their guidelines for conducting systematic reviews, but these sections raise the debates rather than offer guidance (Kahn *et al.*, 2001). There are also several very recent texts on systematic reviews which have built in sections on 'qualitative' research (Jackson and Waters, 2005; Jackson *et al.*, 2005; Petticrew and Roberts, 2005; Popay *et al.*, 2006). Again, although useful, these texts mainly raise the debates rather than offer guidance, pointing out that there is not yet an authoritative body of knowledge to underpin such

guidance in contrast to the body of knowledge which has been developed to underpin systematic reviews of trials (Popay *et al.*, 2006).

This thesis has contributed to, but has not resolved, some of the debates that recent texts raise about assessing the quality of 'qualitative' research in systematic reviews. Its findings do, however, offer a number of concrete suggestions for those who want to undertake a systematic review which includes 'qualitative' and other types of research about intervention processes and people's perspectives and experiences.

In the first instance, it is important to clarify the question for the review and the purpose of the synthesis and to think about how different types of research might answer the review question and provide material consistent with the purpose of the synthesis. Given the importance of the distinction between 'interpretive' and 'aggregative' syntheses for the findings of this thesis I consider it to be worthwhile thinking about the review question and the purpose of the synthesis separately. The findings of this thesis suggest that it is possible to have at least two different purposes with the same review question. For example, with a question of the form 'what are people's perspectives and experiences on x?', the purpose of the synthesis could be to bring together and list aspects of people's perspectives (a synthesis with an aggregative purpose) or to develop concepts, explanations and theory (a synthesis with an interpretive purpose). This first step ought to be a crucial element for planning how quality will be assessed.

Once question and purpose have been clarified, instead of applying just any of the existing tools for assessing the quality of 'qualitative' research, reviewers need to choose a tool which appears to be consistent with their review question and/or purpose of their synthesis. For example, tools by Long and Godfrey (2004), Popay

et al. (1998), Spencer *et al.* (2003) and Whittmore *et al.* (2001) would match quite well with a review question about people's perspectives and experiences. These tools all offer detailed items which would help a reviewer to assess studies according to how well they attend to generating understanding of the world from the point of view of the people studied and diversity in perspective. If the purpose of the synthesis is to generate concepts and theory, then Corbin and Strauss (1990), Cesario *et al.* (2002) and Sandelowski and Barroso (2002) all offer tools that would match this purpose. In either case, reviewers must be prepared to engage with the findings of studies as well as the methods sections of study reports to fully assess quality. The just mentioned tools all offer detailed items which would help a reviewer to assess the quality of the concepts and theory generated by the study. From a practical point of view, reviewers might want to consider using those tools which were evaluated as the most useful tools in this thesis (Campbell *et al.*, 2003; Sandelowski and Barroso, 2002). Although not as detailed as the tools just mentioned, Campbell *et al.* (2003) do attend to both diversity in perspective and quality in the generation of concepts and theory.

The tool by Sandelowski and Barroso (2002) might be a useful starting point for those who want a tool that places emphasis on helping a reviewer to assess whether a study contains 'fatal' as opposed to minor, insignificant flaws. For each item in this tool, reviewers are asked to think about whether the quality issue under consideration is relevant given the appraisal context. Reviewers may have to adapt existing tools, or indeed design a new one, if they want quality assessment items to address in a more direct way whether the findings and conclusions of the study can be trusted to answer the review question given the way the study has been designed and conducted. Another consideration for any new tool, or adaptations to existing tools, is the form of study findings. As noted earlier there may be different 'fatal flaws' depending on the form of study findings.

The findings of this thesis on the relationship between the quality of 'qualitative' research and systematic review results provide a tentative basis for excluding studies from syntheses on intervention processes and people's perspectives and experiences. However, because of the limitations of this thesis, reviewers should proceed with caution. It would be helpful if reviewers could reflect on and share with others what made a good or bad study for their particular synthesis and conduct their synthesis in such a way that they or others could go back to assess the impact of high and low quality studies on synthesis results.

b) Implications for future methodological work

The implications of the findings of this thesis for how to include and quality assess 'qualitative' research in systematic reviews suggests plenty of scope for more methodological work. For example, any new tools developed which link quality assessment to review questions should be evaluated and revised accordingly. From the survey and evaluation of existing tools reported in chapter five, there is also scope for new tools to consider in more detail the issue of sample and sampling. Items in existing tools on sample and sampling were in a minority, but a recent proposal puts generalisability at the top of a new hierarchy of evidence in 'qualitative' research (Daly *et al.*, 2007). In this four-level hierarchy, single case studies appear at the bottom, followed by studies with little detailed analysis but lots of descriptive quotations. The next two levels of evidence are from studies of a more conceptual nature, but only those which are based on an appropriately diverse sample with all data accounted for provide the highest level of evidence in the hierarchy.

Another major gap revealed by the survey and evaluation of existing tools for assessing the quality of 'qualitative' research was the lack of an empirical basis for

the various strategies advocated by the tools for increasing rigour (e.g. the use of more than one researcher to conduct data analysis, respondent validation). This suggests that new research is needed to assess the value of using these strategies when conducting 'qualitative' research. However, in the course of undertaking this thesis I came across a number of studies which had tested some of these strategies (e.g. Armstrong *et al.*, 1997; Greene *et al.*, 1998; Hinds *et al.*, 1990). It seems highly plausible that there are more studies out there as I did not search systematically for such studies. If this is the case then a systematic attempt to identify and summarise these studies would be extremely valuable before mounting new primary research. Such a review would form an important part of a strategy to establish an evidence base to underpin approaches to assessing the quality of 'qualitative' research in systematic reviews and beyond.

The third new methodological study in my thesis was a rare attempt to examine the relationship between the quality of 'qualitative' research and the findings of systematic reviews. As Dixon-Woods *et al.* (2006) also argue, the worth and achievability of conducting sensitivity analyses – in which the effect on synthesis of including and excluding findings from studies of differing quality - should be an important focus of future work. Such work could build on some of the lessons learned from my attempt to study this issue. Reflecting on the problems that I faced suggested two new studies. The first, which would avoid the rather crude synthesis contribution score that I used to examine the relationship between study quality and synthesis results, would be to sort studies relevant to a given synthesis according to high and low quality and conduct separate syntheses on each set of studies. The results of the syntheses of low quality studies could then be compared to the results of the synthesis of high quality studies. The second, which would avoid the possibility that reviewers may unconsciously downplay the findings of lower quality

studies in a synthesis, would involve a replication of study three with the use of different researchers to assess quality and conduct the synthesis.

A final implication from the findings of this thesis concerns who should *conduct* new methodological work. As outlined earlier, the methodological work in this thesis has been conducted by a researcher with a particular perspective on 'qualitative' and 'quantitative' research. At several points in this thesis I have reflected on the potential value of engaging a researcher with a 'qualitative' identity who considers 'qualitative' research to be a completely different kind of inquiry to 'quantitative' research to add a different and challenging perspective. Whilst other research teams have highlighted the value of their collective 'qualitative' expertise (Attree and Milton, 2006; Campbell *et al.*, 2003), I would argue that research teams should aim for a balance between 'qualitative' researchers, 'quantitative' researchers and those researchers who are happiest resisting traditional methodological boundaries.

c) Implications for primary research

The implications outlined above for assessing the quality of 'qualitative' and other types of research on intervention processes and people's perspectives and experiences in systematic reviews rest on the assumption that these studies are well reported. However, the studies used for the methodological work in study three did not meet this assumption (see also Harden et al., 2001a; 2004; Oakley, 2004; Rees et al., 2006; Shepherd et al., 2006). Reporting quality was most problematic with respect to presenting an adequate description of the sample and how it was recruited and reporting data analysis methods adequately. It was often difficult to tell from research reports who was included and excluded from the sample. Whilst reporting on participants sex and age was generally good, details on participants socio-economic background were sketchy, and the ethnicity of children and young

people was largely not reported. When detail was given on data analysis, this was generally limited to statements such as 'data were analysed thematically' or 'common themes were identified from interview transcripts'. Few studies provided descriptions of the development of themes and how data were allocated to codes. It is not certain whether this reflects a lack of skill in conducting qualitative analysis, an acceptance that the procedures used are non-codifiable, or the restrictions on word limits from journal editors. Whatever the reason, the lack of detail is a problem for those who want to track the link between data, interpretations and conclusions.

As mentioned earlier this state affairs does not appear to be unique to 'qualitative' research in health promotion and public health (Long and Godfrey, 2004; O'Conner *et al.*, 2001; Oakley, 2003). Now may be the time to launch a set of guidelines for reporting quality in 'qualitative' research which journal editors could subscribe to. Such guidelines have been produced for several research types in healthcare journals – the CONSORT statement for reporting the results of randomised controlled trials (Begg *et al.*, 1996); the TREND statement for non-randomised controlled trials (Des Jarlais *et al.*, 2004); and the STARD statement for diagnostic studies (Bossuyt *et al.*, 2003) - and there is some evidence that these guidelines have driven up the quality of reporting (Graf *et al.*, 2002; Moher *et al.*, 2001). Although some journals in healthcare, for example the *British Medical Journals*, do have guidelines for reporting 'qualitative' research, it is not currently usual practice in social science journals to offer guidance to authors to ensure that they have fully reported on their methods.

Another implication for primary research from this thesis is to move beyond, or even abandon, the terms 'qualitative' and 'quantitative'. This thesis found that the terms 'qualitative' and 'quantitative' were not particularly illuminating descriptions of the studies of interest. Moreover, in order to assess the relationship between the quality

of 'qualitative' research and the findings of reviews in an adequate way, detailed description of the component parts of studies was required, rather than a simple characterisation of a study as either 'qualitative' or 'quantitative'. These detailed descriptions - which broke studies down into aims, data collection methods, data analysis methods, and type of findings - revealed how differences between studies transcended traditional 'qualitative' and 'quantitative' boundaries. For example, some of the studies displaying characteristics usually considered to be 'quantitative' (e.g. specification of variables in advance and the use of fixed response options), also possessed characteristics usually invoked when describing the special or distinctive features of 'qualitative' research (e.g. highly appropriate methods for studying people's perspectives and experiences and findings which displayed conceptual depth and explanatory power). Attention to the processes and procedures actually employed in research may reveal new ways to describe similarities and differences within research.

8.5 Conclusion

By building on the lessons and gaps within the literatures on assessing the quality of research, this thesis has advanced knowledge about how to include 'qualitative' research in systematic reviews and how to make assessments of the quality of 'qualitative' research in a number of ways. It has brought together and summarised the content of existing tools to assess the quality of 'qualitative' research and, in contrast to previous work, revealed that tools differ according to their emphasis on the methods or findings of 'qualitative' research rather than on epistemological or philosophical grounds. A number of limitations were identified amongst existing tools, not least of which was the finding that tools were not very good at distinguishing between high and low quality research, despite this being one of the stated purposes of tools. I have therefore argued that tools for assessing the quality

of 'qualitative' research are another site where the social science paradigm wars have been played out. I have also highlighted that the actual content of quality assessment tools undermines the argument that a distinctive approach is needed for the assessment of 'qualitative' research. Nearly all items in the tools could be applied to 'quantitative' as well as 'qualitative' research. The ability to step outside the paradigm wars was a crucial influential factor identified by this thesis for the development of a innovative new tool to assess the quality of 'qualitative' research that was, in contrast to previous tools, driven by research questions rather than methods or epistemological position. Despite the advances made by this new tool, the findings of this thesis also suggest that another stage of development is required in order to ensure that the tool a) helps reviewers to assess the quality of findings as well as methods, and b) caters for the diversity in forms of findings produced by 'qualitative' research.

The findings of this thesis have helped to shed light on whether the 'quantitative' model of systematic review fits 'qualitative' research. I have argued that the 'quantitative' model can be useful for conducting systematic reviews of 'qualitative' research if 'qualitative' research is seen as a complementary, rather than competing, type of inquiry in relation to 'quantitative' research. Although existing tools for assessing the quality of 'qualitative' research are ill equipped to do so, assessing the quality of 'qualitative' research for systematic reviews should consider, in relation to the research question, the extent to which study methods have minimised the introduction of bias and error into study findings – the same guiding principles that are used to assess the quality of trials in systematic reviews of effectiveness. The findings of this thesis also support the exclusion from reviews of poor quality studies suggested by the 'quantitative' model of systematic reviews. For example, in contrast to the findings of high quality studies, low quality studies offered only a partial picture of people's perspectives and experiences. However, this thesis also

found that to fully assess the quality of 'qualitative' research, reviewers need to engage with study findings as well as methods, and this represents an important challenge to the 'quantitative' model of systematic reviews.

Finally, this thesis represents a challenge to one of the most fundamental distinctions in social science research methods – that between 'qualitative' and 'quantitative' research. It would have been very difficult to make the methodological advances in this thesis without stepping outside these traditional boundaries. The starting point for this thesis was the need to develop systematic reviews to answer questions about need, context and process and, in the 'real world', the kinds of studies that get conducted do not neatly fall into 'qualitative' and 'quantitative' categories. I have therefore argued that the 'qualitative' and 'quantitative' categories should not be relied upon as a precise way to describe research. Because this thesis has uncovered evidence that the social science paradigm wars are being re-created in the literature on research synthesis I have argued for a reframing of the debate about whether or not the 'quantitative' model of systematic reviews fits 'qualitative' research to one about whether or not the model of systematic reviews developed to answer questions about the effects of interventions fits systematic reviews to answer questions about intervention processes and people's perspectives and experiences.

References

Abraham N, Moayyedi P, Daniels B, Veldhuyzen Van Santen S (2004) The methodological quality of trials affects estimates of treatment efficacy in functional (non-ulcer) dyspepsia. *Alimentary Pharmacology and Therapeutics* **19**: 631-641.

Ackroyd B (1996) The quality of qualitative methods: qualitative or quality methodology for organisation studies. *Organisation* **3**: 439-451.

Aggleton P, Moddy D (1992) Monitoring and evaluating HIV/AIDS health education and health promotion. In: Aggleton P, Young A, Moddy D, Kapila M, Pye M (eds) *Does it Work: Perspectives on the Evaluation of HIV/AIDS Health Promotion*. London: Health Education Authority.

Aggleton P, McClean C, Taylor-Laybourn A, Waller D, Warwick I, Woodhead D, Youdell D (1995) *Young Men Speaking Out*. London: Health Education Authority.

Altheide D, Johnson J (1998) Criteria for assessing interpretive validity in qualitative research. In: Denzin N, Lincoln Y (Eds.) *Handbook of Qualitative Research*. London: Sage Publications.

Angen M (2000) Evaluating interpretive inquiry: reviewing the validity debate and opening the dialogue. *Qualitative Health Research* **10**:378-395.

Arai L (2003) Low expectations, sexual attitudes and knowledge: explaining teenage pregnancy and fertility in English communities. Insights from qualitative research. *Sociological Review* **51**: 199-217.

Arai L, Roen K, Roberts H, Popay J (2005) It might work in Oklahoma but will it work in Oakhampto? Context and implementation in the effectiveness literature on domestic smoke detectors. *Injury Prevention* **11**: 148-151.

Armstrong C, Hill M, Secker J (1998) *Listening to Children*. London: Mental Health Foundation.

Armstrong D, Gosling A, Weinman J, Marteau T (1997) The place of inter-rater reliability in qualitative research: an empirical study. *Sociology* **31**: 597-606.

Attree P (2004) Growing up in disadvantage: a systematic review of the qualitative literature. *Child: Care, Health and Development* **30**: 679-689.

Attree P, Milton B (2006) "Critically appraising qualitative research for systematic reviews: defusing the methodological cluster bombs" *Evidence and Policy* **18**: 109-126.

Bailey R (2001) Overcoming veriphobia – learning to love truth again. *British Journal of Educational Studies* **49**: 159-172.

Balding J, Gimber P, Regis D, Wise A (1997) A quarter of year 7 young men want to cycle to school. *Education and Health* **15**: 49-52.

Balding J, Regis D, Wise A (1998) *No Worries? Young People and Mental Health: A study of the worries and concerns that affect young teenagers in our society*. Exeter: School Health Education Unit.

Barbour R (1999) The case for combining qualitative and quantitative approaches in health services research. *Journal of Health Services Research and Policy* **4**: 39-43.

Barbour R (2001) Checklists for improving rigour in qualitative research: a case of the tail wagging the dog? *British Medical Journal* **322**: 1115-1117.

Barbour R, Barbour M (2003) Evaluating and synthesizing qualitative research: the need to develop a distinctive approach. *Journal of Evaluation in Clinical Practice* **9**: 179-186.

Baxter I, Schroder M, Bower J (2000) Children's perceptions of and preferences for vegetables in the West of Scotland: the role of demographic factors. *Journal of Sensory Studies* **15**: 361-381.

Beck (1993) Qualitative research: the evaluation of its credibility, fittingness and auditability. *Western Journal of Nursing Research* **15**: 263-266.

Becker H (1979) Do photographs tell the truth? In: Cook TD, Reichardt C (eds) *Qualitative and Quantitative Methods in Evaluation Research*. Beverly Hills, California: Sage Publications.

Begg C, Cho M, Eastwood S, Horton R, Moher D, Olkin I, Pitkin R, Rennie D, Schulz K, Simel D, Stroup D (1996) Improving the quality of reporting randomised controlled trials: The CONSORT statement. *Journal of the American Medical Association* **276**: 637-639.

Biemer P, Lyberg L (2003) *Introduction to Survey Quality*. Chichester: Wiley.

Birtwistle GE, Brodie DA (1991) Children's attitudes towards activity and perceptions of physical education. *Health Education Research* **6**: 465-478.

Black N (1997) A national strategy for research and development: lessons from England. *Annual Review of Public Health* **18**: 485-505.

Blaxter M (1996) Criteria for the evaluation of qualitative research papers. *Medical Sociology News* **22**: 68-71.

Bloor M (1997) Techniques of validation in qualitative research: a critical commentary. In Miller G, Dingwall R (Eds.) *Context and Methods in Qualitative Research*. London: Sage Publications.

Blunkett D (2000) Influence or irrelevance: can social science improve government? *Research Intelligence* **71**: 12-21.

Bonell C (1999) Evidence-based nursing: a stereotyped view of quantitative and experimental research could work against professional autonomy and authority. *Journal of Advanced Nursing* **30**: 18-23.

Booth A (2001) Cochrane of cock-eyed? How should we conduct systematic reviews of qualitative research? Paper presented at the *Qualitative Evidence-Based Practice Conference* University of Coventry 14 to 16th May 2001.

Borreani C, Miccinesi G, Brunelli C, Lina M (2004) An increasing number of qualitative research papers in oncology and palliative care: does it mean a thorough development of the methodology of research? *Health and Quality of Life Outcomes* **2**: 7 (http://www.hqlo.com/content/2/1/7).

Bossuyt (2003) P, Reitsma J, Bruns D, Gatsonis C, Glasziou P, Irwing L, Lijmer J, Moher D, Rennie D, deVet H, Standard for Reporting of Diagnostic Accuracy (2003) Towards complete and accurate reporting of diagnostic studies: The STARD initiative. *British Medical Journal* **326**: 41-44.

Boulton M, Fitzpatrick R, Swinburn C (1996) Qualitative research in healthcare II: a structured review and evaluation of studies. *Journal of Evaluation in Clinical Practice* **2**:171-179.

Boulton M, Fitzpatrick R (1997) Evaluating qualitative research. *Evidence-based Health Policy and Management* **December**: 83-85.

Bowen C (1997) School survey highlights teenage problem area. *Nursing Times* **93**: 54-55.

Breese J, Ra'el K, Grant G (2000) No place like home: a qualitative investigation of social support and its effects on recidivism. *Sociological Practice* **2**: 1-21.

Britten N, Jones R, Murphy E, Stacy R (1995) Qualitative research methods in general practice and primary care. *Family Practice* **12**:104-114.

Britten N, Campbell R, Pope C, Donovan J, Morgan M, Pill R (2002) Using meta ethnography to synthesise qualitative research: a worked example. *Journal of Health Services Research and Policy* **7**:209-215.

Britton A, McKee M, Black N, McPherson K, Sanderson G, Bain C (1998) Choosing between randomised and non-randomised studies: a systematic review. *Health Technology Assessment Reports* **2**: 1-124.

Brunton G, Harden A, Rees R, Kavanagh J, Oliver S, Oakley A (2003) *Children and Physical Activity: A systematic review of barriers and facilitators*. London: EPPI-Centre, Social Science Research Unit, Institute of Education, University of London.

Brunton V, Thomas J, Harden A, Rees R, Kavanagh J, Oliver S, Shepherd J, Oakley A (2005) Promoting physical activity amongst children outside of physical education classes: a systematic review integrating intervention studies and qualitative studies. *Health Education Journal* 64: 323-338.

Bryman A (1988) *Quantity and Quality in Social Research*. London: Unwin Hyman.

Burrows C, Eves F, Cooper D (1999) Children's perceptions of exercise: are children mini-adults? *Health Education* **99**: 61-69.

Bushman B (1994) Vote counting procedures in meta-analysis. In: Cooper H, Hedges L (eds) *The Handbook of Research Synthesis*. New York: Russell Sage Foundation.

Cabinet Office (1999) *Modernizing Government*. London: Cabinet Office.

Campbell Collaboration (2001) *Campbell Systematic Reviews: Guidelines for the preparation of review protocols (version 1.0)* http://www.campbellcollaboration.org/c2_protocol_guidelines%20doc.pdf

Campbell DT, Stanley J (1963) *Experimental and Quasi-Experimental Designs for Research*. Boston: Houghton, Mifflin Company.

Campbell DT, Stanley J (1966) *Experimental and Quasi-experimental Designs for Research*. Boston: Houghton Mifflin Company.

Campbell DT (1986) Relabeling internal and external validity for applied social scientists. *New Directions for Program Evaluation* **31**: 67-77.

Campbell R, Pound P, Pope C, Britten N, Pill R, Morgand M, Donovan J (2003) Evaluating meta-ethnography: a synthesis of qualitative research on lat experiences of diabetes and diabetes care. *Social Science and Medicine* **56**: 671-684

Cesario S, Morin K, Santa-Donato A (2002) Evaluating the level of evidence of qualitative research. *Journal of Obstetric, Gynecologic and Neonatal Nursing* **31**: 708-714.

Chaiken M (1990) Evaluation of girls clubs of America's friendly PEERsuasion program. In: Watson R (ed) *Drug and Alcohol Abuse Prevention*. London: The Humana Press Inc. pp 95-132.

Chalmers I (2001) Comparing like with like: some historical milestones in the evolution of methods to create unbiased comparison groups in therapeutic experiments. *International Journal of Epidemiology* **30**:1156-1164.

Chalmers I (2003) Trying to do more good than harm in policy and practice: the role of rigorous, transparent and up-to-date evaluations. *The ANNALS of the American Academy of Political and Social Science* **589**: 22-40.

Chalmers I, Haynes B (1994) Systematic reviews: reporting, updating and correcting systematic reviews of the effects of healthcare. *British Medical Journal* **309**: 862-865.

Chalmers I, Hedges L, Cooper H (2002) A brief history of research synthesis. *Evaluation & the Health Professions* **25**: 12-37.

Chalmers TC, Smith H Jr, Blackburn B, Silverman B, Schroeder B, Reitman D, Ambroz A (1981) A method for assessing the quality of a randomized control trial. *Controlled Clinical Trials* **2**: 31-49.

Chalmers TC, Celano P, Sacks H, Smith H Jr (1983) Bias in treatment assignment in controlled clinical trials. *New England Medical Journal* **309**:1358-61.

Chambers J, Cleveland W, Kleiner B, Tukey P (1983). *Graphical Methods for Data Analysis*. Belmont, CA: Wadsworth.

Chapple A, Rodgers A (1998) Explicit guidelines for qualitative research: a step in the right direction, a defence of the 'soft' options, or a form of sociological imperialism? *Family Practice* **15**: 556-561.

Charleston S, Oakley A, Johnson A, Stephenson J, Brodala A, Fenton K, Petruckevitch A (1996) *Report on a Pilot Study for a Randomised Controlled Trial of Peer-led Sex Education*. London: Social Science Research Unit.

Coakley J, White A (1992) Making decisions - gender and sport participation among British adolescents. *Sociology of Sport Journal* **9**: 20-35.

Cobb A, Hagemaster J (1987) Ten criteria for evaluating qualitative research

proposals. *Journal of Nursing Education* **26**:138-143.

Cook TD, Campbell DT (1979) *Quasi-Experimentation: Design and analysis issues for field settings*. Boston: Houghton Mifflin Company.

Cook TD, Reichardt C (1979*) Qualitative and Quantitative Methods in Evaluation Research*. Beverly Hills, California: Sage Publications.

Cooper H, Hedges L (1994) *The Handbook of Research Synthesis*. New York: Russell Sage Foundation.

Corbin and Strauss (1990) Grounded theory research: procedures, canons, and evaluation criteria. *Qualitative Sociology* **13**: 3-21

Creswell J (1995) *Research Design: Qualitative and quantitative techniques*. Thousand Oaks, CA: Sage Pulications.

Creswell J, Miller D (2000) Determining validity in qualitative inquiry. *Theory into Practice* **39**: 124-130.

Critical Appraisal Skills Programme (1998) *Ten Questions To Help You Make Sense Of Qualitative Research.* Milton Keynes: Milton Keynes Primary Care Trust.

Critical Appraisal Skills Programme (2002) *Ten Questions To Help You Make Sense Of Qualitative Research.* Milton Keynes: Milton Keynes Primary Care Trust.

Crowley P (1996) Prophylactic corticosteroids for preterm birth. *The Cochrane Database of Systematic Reviews* Issue 1. Art. No.: CD000065. DOI: 10.1002/ 14651858.CD000065

Daly A, Willis K, Small R, Green J, Welch N, Kealy M, Hughes E (2007) Hierarchy of evidence for assessing qualitative health research. *Journal of Clinical Epidemiology* **60**: 43-49.

Darbyshire P (1997) Qualitative research: Is it becoming the new orthodoxy? *Nursing Inquiry* **4**: 1-2.

Davies JK, MacDonald G (1998) *Quality, Evidence and Effectiveness in Health Promotion*. London: Routledge.

Davies P (1999) What is evidence-based education? *British Journal of Educational Studies* **47**: 108-121.

Davis A, Jones L (1996) Environmental constraints on health: listening to children's views. *Health Education Journal* **55**: 363-374.

Day S, Altman D (2000) Blinding in clinical trials and other studies. *British Medical Journal* **321**: 504.

Deeks J, Altman D, Bradburn M (2001) Statistical methods for examining heterogeneity and combining results from several studies in meta-analysis. In Egger M, Davey-Smith G, Altman D (eds) *Systematic Reviews in Health Care: Meta-analysis in context*. London: BMJ Publishing.

Deeks J, Dinnes J, D'Amico R, Sowden A, Sakarovitch C, Song F, Petticrew M, Altman D (2003) Evaluating non-randomised intervention studies. *Health Technology Assessment Reports* **7**(27).

Dennison CM, Shepherd R (1995) Adolescent food choice: an application of the theory of planned behaviour. *Journal of Human Nutrition and Dietetics* **8**: 9-23.

Denzin N, Lincoln Y (2005) *Handbook of Qualitative Research.* (3rd Edition). London: Sage Publications.

Department of Health (1998) *A First Class Service: Quality in the new NHS.* London: The Stationery Office.

Department of Health (1999a) *Saving Lives: Our healthier nation*. London: The Stationery Office.

Department of Health (1999b) *Patient and Public Involvement in the New NHS.* London: The Stationery Office.

Department of Health (2001a) *A Research and Development Strategy for Public Health*. London: The Stationery Office.

Department of Health (2001b) *Better Prevention, Better Services, Better Health: The national strategy for sexual health and HIV*. London: The Stationery Office.

Derbyshire J (1996) The feel-bad factor. *Young People Now* **October**: 20-21.

Des Jarlais D, Lyles C, Crepaz N, TREND Group (2004) Improving the reporting quality of non-randomized evaluations of behavioural and public health interventions: The TREND statement. *American Journal of Public Health* **94**: 361-366.

DeVaus D (2002) *Surveys in Social Research*. London: Routledge.

Devers K (1999) How will we know "good" qualitative research when we see it? Beginning the dialogue in health services research. *Health Services Research* **34**:1153-1188.

DiCenso A, Guyatt, G, Willan A, Griffith L (2002) Interventions to reduce unintended teenage pregnancies in adolescents: a systematic review of randomised controlled trials. *British Medical Journal* **324**:1426-1434.

Dixey R, Sahota P, Atwal S, Turner A (2001) Children talking about healthy eating: data from focus groups with 300 9-11-year-olds. *Nutrition Bulletin* **26**: 71-79.

Dixon-Woods M, Fitzpatrick R (2001) Qualitative research in systematic reviews has established a place for itself. *British Medical Journal*. 323:765-766.

Dixon-Woods M, Fitzpatrick R, Roberts K (2001) Including qualitative research in systematic reviews: problems and opportunities. *Journal of Evaluation in Clinical Practice* **7**:125-33.

Dixon-Woods M, Agarwal S, Young B, Jones DR, Sutton A (2004a) *Integrative Approaches to Qualitative and Quantitative Evidence*. London: Health Development Agency.

Dixon-Woods M, Shaw R, Agarwal S, Smith J (2004b) The problem of appraising qualitative research. *Quality and Safety in Health Care* 13:223-225.

Dixon-Woods M, Agarwal S, Jones D, Young B, Sutton A (2005) Synthesising qualitative and quantitative evidence: a review of possible methods. *Journal of Health Services Research and Policy* 10: 45-53.

Dixon-Woods M, Bonas S, Booth A, Jones D, Miller T, Sutton A, Shaw R, Smith J, Young B (2006) How can systematic reviews incorporate qualitative research? A critical perspective. *Qualitative Research* **6**: 27-44.

Docherty B (2002) A survey of sexual activity in HIV-positive gay men. *NT Research* **7**: 139-151.

Dodds J, Nardone A, Mercey D, Johnson A (2000) Increase in high risk sexual behaviour among homosexual men, London 1996-8: Cross sectional, questionnaire study. *British Medical Journal* **320**: 1510-1511.

Douglas D (2003) Reflections on research supervision: a grounded theory case of reflective practice. *Research in Post-Compulsory Education* **2**: 213-230.

Drisko (1997) Strengthening qualitative studies and reports: standards to promote academic integrity. *Journal of Social Work Education* **33**: 185-195.

Dyson A, Howes A, Roberts B, (2002) A systematic review of the effectiveness of school-level actions for promoting participation by all students (EPPI-Centre Review, version 1.1). In: *Research Evidence in Education Library*. London: EPPI-Centre, Social Science Research Unit, Institute of Education.

Eakin J, Mykhalovskiy E (2003) Reframing the evaluation of qualitative health research: reflections on a review of appraisal guidelines in the health sciences. *Journal of Evaluation in Clinical Practice* **9**: 187-194.

Edwards J, Hartwell H (2002) Fruit and vegetables: attitudes and knowledge of primary school children. *Journal of Human Nutrition and Dietectics* **15**: 365-374.

Egger M, Davey-Smith G (2001) Principles of and procedures for systematic reviews. In: Egger M, Davey-Smith G, Altman D (eds*) Systematic Reviews in Health Care: Meta-analysis in context*. London: BMJ Publishing.

Egger M, Davey-Smith G, Altman D (2001) *Systematic Reviews in Health Care: Meta-analysis in context*. London: BMJ Publishing.

Elbourne D, Oakley A, Gough D (2001) EPPI-Centre reviews will aim to disseminate systematic reviews in education (letter). *British Medical Journal* **323**:1252.

Elder and Miller (1995) Reading and evaluating qualitative research studies. *The Journal of Family Practice* **41**: 279-285.

Elek-Fisk E, Raymond L, Wortman P (2000) Validity applied to meta-analysis and research synthesis. In Bickman L (ed.) *Validity and Social Experimentation: Donald Campbell's legacy*. Thousand Oaks, California: Sage Publications.

Elliot R, Fischer C, Rennie D (1999) Evolving guidelines for publication of qualitative research studies in psychology and related fields. *British Journal for Clinical Psychology* **38**: 215-229.

EPPI-Centre (2002) *Guidelines for Extracting Data and Quality Assessing Primary Studies in Educational Research (version 0.97)*. London: EPPI-Centre, Social Science Research Unit.

EPPI-Centre (2006) *EPPI-Centre Methods for Conducting Systematic Reviews.* London: EPPI-Centre, Social Science Research Unit, Institute of Education, University of London.

Evans J, Benefield P (2001) Systematic reviews of educational research: does the medical model fit? *British Educational Research Journal* **27**: 527-541.

Felix-Oritz M, Salazar M, Gonzalez J, Sorensen J, Plock D (2003) A qualitative evaluation of an assisted self-help group for drug addicted clients in a structured outpatient treatment setting. *Community Mental Health Journal* **36**:339-350.

Fenton E, Harvey J, Griffiths F, Wild A, Sturt J (2001) Reflections from organization science on the development of primary health care research networks. *Family Practice* **18**: 540-544.

Fife Healthcare NHS Trust (1996) *Peer Education HIV/AIDS evaluation report, 1, 2 and 3.* Fife: Fife Healthcare NHS Trust.

Fisher M (2005) Knowledge production for social welfare in P Sommerfeld (Ed) *Evidence-Based Social Work: Towards a new professionalism?* Bern: Peter Lang.

Fitzpatrick R, Boulton M (1994) Qualitative methods for assessing healthcare. *Quality in Health Care* **3**: 107-113.

Forchuck C, Roberts J (1993) How to critique qualitative research articles. *Canadian Journal of Nursing Research* **25**: 47-56.

Fox J, Walker B, Kushner S (1993) *"It's not a bed of roses": Young mothers' education project evaluation report*. Norwich: Centre for Applied Research in Education, University of East Anglia.

Francis B, Skelton C, Archer L (2002). A systematic review of classroom strategies for reducing stereotypical gender constructions among girls and boys in mixed–sex UK primary schools (EPPI-Centre Review, version 1.1). In: *Research Evidence in Education Library*. London: EPPI-Centre, Social Science Research Unit, Institute of Education.

Frankham J (1993) *AIDS Peer Education Project Evaluation Report*. Norwich: Centre for Applied Research in Education, University of East Anglia

Friedli L, Scherzer A (1996) *Positive Steps: Mental health and young people: Attitudes and awareness among 11-24 year olds*. London: Health Education Authority.

Fuqua J, Stokols D, Gress J, Phillips K, Harvey R (2004) Transdisciplinary collaboration as a basis for enhancing the science and prevention of substance use and "abuse". *Substance Use and Misuse* **39**: 1457-1514.

Furlong J, Oancea A (2005) *Assessing Quality in Applied and Practice-based Educational Research: A framework for discussion*. Oxford: Oxford University Department of Educational Studies.

Gallagher M, Millar R, Hargie O, Ellis R (1992) The personal and social worries of adolescents in Northern Ireland: results of a survey. *British Journal of Guidance and Counselling* **20**: 274-290.

Gallagher M, Millar R (1996) A survey of adolescent worry in Northern Ireland. P*astoral Care in Education* **14**: 26-32.

Garcia J, Sinclair J, Dickson K, Thomas J, Brunton J, Tidd M, the PSHE Review Group (2006) Conflict resolution, peer mediation and young people's relationships. Technical Report. In: *Research Evidence in Education Library*. London: EPPI-Centre, Social Science Research Unit, Institute of Education, University of London.

Garrett D, Hodkinson P (1998) Can there be criteria for selecting research criteria? A hermeneutical analysis of an inescapable dilemma. *Qualitative Inquiry* **4**: 515-539.

Geertz C (1983) *Local Knowledge: Further essays in interpretive anthropology*. New York: Basic Books.

Gentle P, Caves R, Armstrong N, Balding J, Kirby B (1994) High and low exercisers among 14- and 15-year-old children. *Journal of Public Health Medicine* **16**: 186-194.

Giacomini M, Cook D (2000a) Users' guides to the medical literature XXIII: qualitative research in health care A. (Are the results of the study valid?). *JAMA* **284**: 357-362.

Giacomini M, Cook D (2000b) Users' guides to the medical literature XXIII: qualitative research in health care B. (What are the results and how do they help me care for my patients). *JAMA* **284**: 478-482.

Gibson E, Wardle J, Watts C (1998) Fruit and vegetable consumption, nutrition knowledge and beliefs in mothers and children. *Appetite* **31**: 205-228.

Gilner J (1994) Reviewing qualitative research: proposed criteria for fairness and rigour. *Occupational Therapy Journal of Research* **14**: 78-90.

Glass G (1976) Primary, secondary and meta-analysis of research. *Educational Researcher* **5**: 3-8.
Glass G (1978) Reply to Masfield and Busse. *Educational Researcher* **7**:3.

Gordon J, Grant G (eds) (1997) *How We Feel: An insight into the emotional world of teenagers*. London: Jessica Kingsley Publishers.

Gough D (2004) Systematic research synthesis. In Thomas G, Pring R (Eds): *Evidence-based Practice in Education.* Buckingham: Open University Press.

Gough D (forthcoming) Weight of evidence: a framework for the appraisal of the quality and relevance of evidence. In J. Furlong, A. Oancea (Eds.) *Applied and Practice-based Research* (Special Edition of Research Papers in Education), Summer 2007.

Gough D, Elbourne D (2002) Systematic research synthesis to inform policy, practice and democratic debate. *Social Policy and Society* **1**:225-236.

Graf J, Doig D, Cook D, Vincent J, Sibbald W (2002) Randomized, controlled clinical trials in sepsis: Has methodological quality improved over time? *Critical Care Medicine* **30**: 461-472.

Greene J, Doughty J, Marquart J, Ray M, Roberts L (1998) Qualitative evaluation audits in practice. *Evaluation Review* **12**: 352-375.

Greenlagh T, Taylor R (1997) How to read a paper: papers that go beyond numbers (qualitative research). *British Medical Journal* **315**: 740-743.

Greenland S (1994) Quality scores are useless and potentially misleading. *American Journal of Epidemiology* **140**:300-302.

Guba E, Lincoln Y (1981) *Effective Evaluation*. San Francisco: Jossey-Bass.

Guy A, Banim M (1991) AIDSBUSTERS: a report on the effectiveness of a young person team in designing and delivering HIV and safer sex training within a Youth Training Scheme. *Youth and Policy* **35**: 1-7.

Guyatt G, DiCenso A, Farewell V, Willan A, Griffith L (2000) Randomized trials versus observational studies in adolescent pregnancy prevention. *Journal of Clinical Epidemiology* **53**: 167-174.

Hammersley M (1992) *What's Wrong With Ethnography?* London: Routledge

Hammersley M (2000a) *Taking Sides in Social Research: Essays on partisanship and bias*. London: Routledge.

Hammersley M (2000b) Varieties of social research: a typology. *International Journal of Social Research Methods* **3**: 221-229.

Hammersley M (2001) On systematic reviews of research literatures: a narrative response to Evans and Benefield. *British Educational Research Journal* **27**: 543-554.

Hammersley M (2002) *Educational Research, Policy-making and Practice*. London: Sage Publications.

Hammersley M, Gomm R (1997) Bias in Social Research. *Sociological Research Online* 2: http://www.socresonline.org.uk/socresonline/2/1/2.html

Harden A (2006) Extending the boundaries of systematic reviews to integrate different types of study: examples of methods developed within reviews on young people's health. In J Popay (Ed) *Moving Beyond Effectiveness in Evidence Synthesis*. London: National Institute for Health and Clinical Excellence

Harden A, Willig C (1998) An exploration of the discursive constructions used in young adults memories and accounts of contraception. *Journal of Health Psychology*, 3, 429-445.

Harden A, Ogden J (1999a) 16-19 year olds beliefs about contraceptive services and the intentions to use contraceptives, *British Journal of Family Planning*, 24, 135-141.

Harden A, Ogden J (1999b) Use and non-use of contraception in 16-19 year olds: A within subjects comparison. *Psychology and Health*, 14, 697-709.

Harden A, Ogden J (1999c) Young women's experiences of abortion services. *Sociology of Health and Illness*, 21, 426-444.

Harden A, Peersman G, Oliver S, Mauthner M, Oakley A. (1999a) A systematic review of the effectiveness of health promotion interventions in the workplace. *Occupational Medicine* **49**:1-9.

Harden A, Weston R, Oakley A (1999b) *A Review of the Effectiveness and Appropriateness of Peer-Delivered Health Promotion for Young People*. London: EPPI-Centre, Social Science Research Unit.

Harden A, Oliver S (2001) Who's listening? Systematically reviewing for ethics and empowerment. In Oliver S, Peersman G (eds.) *Using Research for Effective Health Promotion*. Buckingham: Open University Press.

Harden A, Oakley A, Oliver S (2001a) Peer-delivered health promotion for young people: A systematic review of different study designs. *Health Education Journal* **60**: 339-353.

Harden A, Rees R, Shepherd J, Brunton G, Oliver S, Oakley A (2001b) *Young People and Mental Health: A systematic review of barriers and facilitators.* London: EPPI-Centre, Social Science Research Unit, Institute of Education, University of London.

Harden A, Garcia J, Oliver S, Rees R, Shepherd J, Brunton G, Oakley A (2004) Applying systematic review methods to studies of people's views: an example from public health. *Journal of Epidemiology and Community Health* **58**: 794-800.

Harden A, Thomas J (2005) Methodological issues in combining diverse study types in systematic reviews. *International Journal of Social Research Methods* **8**: 257-271.

Harding S (1991) *Whose Science? Whose Knowledge? Thinking from women's lives*. New York: Cornell University Press.

Hargreaves DH (1996) Teaching as a research-based profession: possibilities and prospects, Teacher Training Agency Annual Lecture 1996. London: Teacher Training Agency.

Harris J (1993) Young people's perceptions of health, fitness and exercise. *Physical Education Research Supplement* **13**: 5-9.

Hart C (1998) *Doing a Literature Review: Releasing the social science research imagination*. London: Sage.

Hart K, Bishop J, Truby H (2002) An investigation into school children's knowledge and awareness of food and nutrition. *Journal of Human Nutrition and Dietetics* **15**: 129-140.

Health Education Authority (1995) *Expectations for the Future: An investigation into the self-esteem of 13 and 14 year old girls and boys*. London: Health Education Authority.

Higgins J, Green S (eds) (2006) Cochrane Handbook for Systematic Reviews of Interventions 4.2.6 [updated September 2006]. http://www.cochrane.org/resources/handbook/hbook.htm (accessed 24th January 2007)

Hillage J, Pearson R, Anderson A, Tamkin P (1998) *Excellence in Research in Schools*. London: Department for Education and Employment.

Hinds P, Scandrett-Hibden S, McAuley L (1990) Further assessment of a methods to estimate the reliability and validity of qualitative research findings. *Journal of Advanced Nursing* **15**: 430-435.

Hoddinott P, Pill R (1997) A review of recently published qualitative research in general practice: more methodological questions than answers? *Family Practice* 14: 313-319.

Hoinville G, Jowell R (1978) *Survey Research Practice*. London: Heinemann Educational.

Hopwood T, Carrington B (1994) Physical education and femininity. *Educational Research* **36**: 237-246.

Hulme D, Toye J (2006) The case for cross-disciplinary social science research on poverty, inequality and well-being. *Journal of Development Studies* **42**: 1085- 1107.

Jackson N, Waters E, for the Guidelines for Systematic Reviews of Health Promotion and Public Health Interventions Taskforce (2005) Guidelines for Systematic reviews of health promotionand public health interventions. Version 1.2. Melbourne: Deakin University.

Jackson N, Waters E, Anderson L, Bailie R, Brunton G, Hawe P, Kristjansson E, Naccarella L, Norris S, Oliver S, Petticrew M, Peinaar E, Popay J, Roberts H, Rogers W, Shepherd J, Sowden A, Thomas H (2005) Criteria for the systematic review of health promotion and public health interventions. *Health Promotion International* **20**: 376-374.

Jadad AR, Moore RA, Carroll D, Jenkinson C, Reynolds DJ, Gavaghan DJ, McQuay HJ (1996) Assessing the quality of reports of ranomized clinical trials: is blinding necessary? *Controlled Clinical Trials* **17**: 1-12.

Johnson A, Mercer C, Erens B, Copas A, McManus S, Wellings K, Fenton K, Korovessis C, Macdowall W, Nanchahal K, Purdon S, Field J (2001) Sexual behaviour in Britain: partnerships, practices, and HIV risk behaviours. *Lancet* **358**: 1835-1842.

Johnson R, Onwuegbuzie A (2004) Mixed methods research: a research paradigm whose time has come? *Educational Researcher* **33**: 14-26.

Jones K (2004) Mission drift in qualitative research, or moving toward a systematic review of qualitative studies, moving back to a more systematic narrative review. *The Qualitative Report* **9**: 95-112.

Juni P, Witschi A, Bloch R, Egger M (1999) The hazards of scoring the quality of clinical trials for meta-analysis. *Journal of the American Medical Association* **282**:1054-1060.

Juni P, Altman D, Egger M (2001a) Assessing the quality of randomised controlled trials. In: Egger M, Davey-Smith G, Altman D (eds*) Systematic Reviews in Health Care: Meta-analysis in context*. London: BMJ Publishing.

Juni P, Altman D, Egger M (2001b) Assessing the quality of controlled clinical trials. *British Medical Journal*. **323**:42-46.

Khan K, ter Riet G, Glanville J, Sowden A, Kleijnen J (2001) *Undertaking Systematic Reviews of Research on Effectiveness:: CRD's guidance for those carrying out of commissioning reviews*. York: Centre for Reviews and Dissemination, University of York.

Katrak P, Bialocerkowski A, Massy-Westropp N, Kumar S, Grimmer K (2004) A systematic review of the content of critical appraisal tools. *BMC Medical Research Methodology* **4**: 22.

Kearney M (2001) Levels and applications of qualitative research evidence. R*esearch in Nursing and Health* **24**:145-153.

Kincey J, Amir Z, Gillespie B, Carleton E, Theaker T (1993) A study of self esteem, motivation and perceived barriers to participation in sport and exercise among secondary school pupils. *Health Education Journal* **52**: 241-245.

Kirk J, Miller M (1984) *Reliability and Validity in Qualitative Research*. London: Sage Publications.

Kleijnen J, Gotzsche P, Kunz R, Oxman A, Chalmers I (1997) So what's so special about randomisation? In: Maynard A, Chalmers I (eds) *Non-random Reflections on Health Services Research*. London: BMJ Publishing Group.

Khun T (1970) *The Structure of Scientific Revolutions (Second Edition)*. Chicago: University of Chicago Press.

Krippendorf K (2004) Content Analysis: An introduction to its methodology. Thousand Oaks: Sage Publications.

Kunz R, Oxman A (1998) The unpredictability paradox: review of empirical comparisons of randomised and non-randomised clinical trials. *British Medical Journal* **317**: 1185-1190.

Kunz R, Vist G, Oxman A (2002) Randomisation to protect against selection bias in healthcare trials. *Cochrane Database of Methodology Reviews* 2002, Issue 4. Art. No.: MR000012. DOI: 10.1002/14651858.MR000012.

Kuzel A, Engel J (2001) Some pragmatic thoughts about evaluating qualitative health research. In: Morse J, Swanson J, Kuzel A (eds) *The Nature of Qualitative Evidence.* Thousand Oaks, California: Sage Publications.

Lather P (1993) Fertile obsession: validity after post structuralism. *Sociological Quarterly* **34**: 673-693.

Lather P (1999) To be of use: the work of reviewing. *Review of Educational Research* **69**: 2-7.

Leach E (1982) *Social Anthropology*. Oxford: Oxford University Press.

LeCompte M, Preissle Goetz J (1982) Problems of reliability and validity in ethnographic research. *Review of Educational Research* **52**: 31-60.

Leininger M (1994) Evaluation criteria and critique of qualitative research studies. In Morse J (Ed) *Critical Issues in Qualitative Research Methods*. Thousand Oaks, California: Sage Publications.

Levacic R, Glatter R (2001) "Really good ideas?" Developing evidence-informed policy and practice in education leadership and management. *Educational Management and Administration* **29**: 5-25.

Light R, Pillemer D (1984) *Summing Up: The science of reviewing research.* Cambridge, MA: Harvard University Press.

Lincoln Y (1995) Emerging criteria for quality in qualitative and interpretive research. *Qualitative Inquiry* **1**: 275-289.

Lincoln Y, Guba E (1985) *Naturalistic Inquiry*. Beverly Hills, California: Sage Publications.

Lipsey M, Wilson D (2001) *Practical Meta-analysis*. Thousand Oaks, California:

Sage.

Livingston G (1999) Beyond watching over established ways: a review as recasting the literature, recasting the lived. *Review of Educational Research* **69**: 9-19.

Long AF, Godfrey M (2004) An evaluation tool to assess the quality of qualitative research studies. *International Journal of Social Research Methodology* **7**: 181-196.

McDougall P (1998) Teenagers and nutrition: assessing levels of knowledge. *Health Visitor* **71**: 62-4.

McGuiness R (1994) *Youth for Health: Peer led education in South East Kent.* Kent: South East Kent Health Authority

McLaughlin J (1986) Reliability and validity issues in school ethnography and qualitative research. *Journal of School Health* **56**: 187-189.

MacLehose R, Reeves B, Harvey I, Sheldon T, Russell, I, Black A (2000) A systematic review of comparisons of effect sizes derived from randomised and non-randomised studies. *Health Technology Assessment Reports* **4**: 1-154.

MacLure M (2005) Clarity bordering on stupidity: where's the quality in systematic review? *Journal of Education Policy* 20: 393-416.

Madill A, Jordan A, Shirley C (2000) Objectivity and reliability in qualitative analysis: realist, contextualist and radical constructionist epistemologies. *British Journal of Psychology* **91**:1-20.

Malterud K (2001) Qualitative research: standards, challenges and guidelines. *The Lancet* **358**: 483-488.

Mark M (2000) Realism, validity, and the experimenting society. In Bickman L (ed.) *Validity and Social Experimentation: Donald Campbell's legacy*. Thousand Oaks, California: Sage Publications.

Mason V (1995) *Young People and Sport in England, 1994: The views of teachers and children*. London: Sports Council.

Massey C, Alpass F, Flett R, Lewis K, Morriss S, Sligo F (2006) Crossing fields: the case of a multi-disciplinary research team. *Qualitative Research* **6**: 131-147.

Mauthner M, Mayall B, Turner S (1993) *Children and Food at Primary School.* London: Social Science Research Unit, Institute of Education, University of London.

Mayring, P (2000). Qualitative Content Analysis [28 paragraphs]. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research [On-line Journal]* **1**: Available at: http://www.qualitative-research.net/fqs-texte/2-00/2-00mayring-e.htm [Accessed 24th April, 2007]

Mays N, Pope C (1995) Rigour and qualitative research. *British Medical Journal* **311**:109-12.

Mays N, Pope C (2000) Qualitative research in health care: Assessing quality in qualitative research. *British Medical Journal* **320**: 50-52.

Mays N, Roberts E, Popay J (2001) Synthesising research evidence. In Fulop N, Allen P, Clarke A, Black N (eds) *Methods for Studying the Delivery and Organisation of Health Services.* London: Routledge.

Mays N, Pope C, Popay J (2005) Systematically reviewing qualitative and quantitative evidence to inform management and policy-making in the health field. *Journal of Health Services Research and Policy* **10**: S1:6-S1:20.

Milburn K (1995) A critical review of peer education with young people with special reference on sexual health. *Health Education Research* **10**: 407-420.

Miles G, Eid S (1997) The dietary habits of young people. *Nursing Times* **93**: 46-8.

Miles M, Huberman A (1994) *Qualitative Data Analysis: An expanded sourcebook.* London: Sage Publications.

Miller B (1993) Femininity, physical activity and the curriculum. In McFee G, Tomlinson A (Eds) *Education, Sport and Leisure: Connections and controversies.* Eastbourne: University of Brighton.

Milton B, Whitehead M, Holland P, Hamilton V (2004) The social and economic consequences of childhood asthma across the lifecourse: a systematic review. *Child: Care, Health and Development* **30**: 711-728

Mitchell K (1997) Encouraging young women to exercise: can teenage magazines play a role? *Health Education Journal* **56**: 264-273.

Moher D, Jadad A, Nichol G, Penman M, Tugwell P, Walsh S (1995) Assessing the quality of randomized controlled trials: an annotated bibliography of scales and checklists. *Controlled Clinical Trials* **16**: 62-73.

Moher D, Pham B, Jones A, Cook DJ, Jadad AR, Moher M, Tugwell P, Klassen TP (1998) Does quality of reports of randomised trials effect estimates of intervention efficacy reported in meta-analyses? *Lancet* **352**: 609-613.

Moher D, Cook DJ, Jadad A, Tugwell P, Moher M, Jones A, Pham B, Klassen T (1999) Assessing the quality of reports of randomised trials: implications for the conduct of meta-analyses. *Health Technology Assessment* **3**:1-55.

Moher D, Jones A, Lepage L for the CONSORT Group (2001) Use of the CONSORT statement and quality of reports of randomized controlled trials: A comparative before and after study. *Journal of the American Medical Association* **285**: 1992-1995.

Muecke M (1994) On the evaluation of ethnographies. In: Morse J (ed.) *Critical Issues in Qualitative Research Methods.* Thousand Oaks, California: Sage Publications.

Mulrow C (1987) The medical review article: state of the science. *Annals of Internal Medicine* **106**:485-8.

Mulrow C (1994) Systematic reviews: rationale for systematic reviews. *British Medical Journal* **309**:597-599.

Mulvihill C, Rivers K, Aggleton P (2000a) *Physical Activity 'At Our Time'.* London: Health Education Authority.

Mulvihill C, Rivers K, Aggleton P (2000b) A qualitative study investigating the views of primary-aged children and parents on physical activity. *Health Education Journal* **59**: 166-179.

Murphy E, Dingwall R, Greatbatch D, Parker S, Watson P (1998) Qualitative research methods in health technology assessment: a review of the literature. *Health Technology Assessment* **2**(16).

Neale R, Otte S, Tilston C (1998) Fruit: comparisons of attitudes, knowledge and preferences of primary school children in England and Germany. *Zeitschrift fur Ernahrungswissenschaft* **37**: 128-130.

Newman R, Smith C, Nutbeam D (1991) Teachers' views of the 'Smoking and Me' project. *Health Education Journal* **50**: 107-110.

Noblitt G, Hare R (1988) *Meta-Ethnography: Synthesizing qualitative studies.* London: Sage.

Noyes J, Popay J, Garner P (2005) What can qualitative research contribute to a Cochrane systematic review of DOT for promoting adherence to tuberculosis treatment? Paper presented at *Qualitative Research and Systematic Reviews workshop,* Continuing Professional Development Centre, University of Oxford, 28-29 June.

Nutbeam D (1999) The challenge to provide evidence in health promotion. *Health Promotion International* **14**: 99-101.

Nutley S, Walter I, Davies H (2003) From knowing to doing: A framework for understanding the evidence-into-practice agenda. *Evaluation* **9**: 142-158.

O'Conner T, Koning F, Meakes E, McLarnon-Sinclair K, Davis K, Loy V (2001) Quantity and rigor of qualitative research in four pastoral counseling journals. *Journal of Pastoral Care* **55**: 271-280.

Oakley A (1998) Experimentation in social science: the case of health promotion. *Social Sciences in Health* **4**: 73-89.

Oakley A (2000) *Experiment in Knowing: Gender and methods in the social sciences*. Cambridge: Polity Press.

Oakley A (2002) Social science and evidence-based everything: the case of education. *Educational Review* **54**: 277-286.

Oakley A (2003) Research evidence, knowledge management and educational practice: early lessons from a systematic approach. *London Review of Education* **1**: 21- 33.

Oakley A (2004) Qualitative research and scientific inquiry. *Australian and New Zealand Journal of Public Health* **28**: 106-108.

Oakley A (2006) Resistances to 'new' technologies of evaluation: education research in the UK: education as a case study. *Evidence and Policy* **2**: 63-87

Oakley A, Fullerton D, Holland J, Arnold S, France-Dawson M, Kelley P, McGrellis S (1995) Sexual health education interventions for young people: a methodological review. *British Medical Journal* **310**: 158-162.

Oakley A, Fullerton D (1996) The lamppost of research: support or illumination? In: Oakley A, Roberts H (eds) *Evaluating Social Interventions: A report on two workshops.* Ilford, Essex: Barnardos.

Oakley A, Oliver S (2001) Looking to the future: policies and opportunities for better health. In: Oliver S and Peersman, G (eds) (2001) *Using Research for Effective Health Promotion.* Buckingham: Open University Press.

Oakley A, Gough D, Oliver S, Thomas J (2005) The politics of evidence and methodology: lessons from the EPPI-Centre. *Evidence and Policy* **1**:1-13.

Oliver S (2001) Making research more useful: integrating different perspectives and different methods. In: Oliver S, Peersman G (eds) *Using Research for Effective Health Promotion*. Buckingham: Open University.

Oliver S, Nicholas A, Oakley A (1996) *Promoting Health After Sifting the Evidence: Workshop report.* London: EPI-Centre Report, Social Science Research Unit, Institute of Education.

Oliver S, Peersman G, Harden A, Oakley A (1999a) Discrepancies in findings from effectiveness reviews: the case of health promotion for older people. *Health Education Journal* **58**:78-90.

Oliver S, Oakley L, Lumley J, Waters E (1999b) Observational and qualitative research: increasing a review's relevance to practitioners and consumers. Paper presented at the *VII Cochrane Colloquium*, Rome 6th to 9th October 1999.

Oliver S, Peersman G (eds) (2001) *Using Research for Effective Health Promotion.* Buckingham: Open University Press

Oliver S, Harden A, Rees R, Shepherd J, Brunton G, Garcia J, Oakley A (2005) An emerging framework for integrating different types of evidence in systematic reviews for public policy. *Evaluation* **11**:428-466.

Oppenheim A (1966) *Questionnaire Design and Attitude Measurement.* London: Heinemann.

Organisation for Economic Co-operation and Development (2003) *New Challenges for Educational Research.* Centre for Educational Research and Innovation, OECD, Paris:OECD Publishing

Orme J (1991) Adolescent young women and exercise: too much of a struggle? Education and Health **9**: 76-80.

Orme J, Starkey F (1999) Peer drug education: the way forward? *Health Education* **1**: 8-16.

Parker I (1989) *The Crisis in Modern Social Psychology and How to End it.* London: Routledge.

Patel G, Maharaj K, Rooprah M, Sandhu S (1999) *Hard to Reach, Hard to Teach: Research into the sexual health of South Asian men who have sex with men.* London: Naz Project.

Paterson B, Thorne S, Canam C, Jillings C (2001) *Meta-study of Qualitative Health Research*. Thousand Oaks, California: Sage.

Patton Quinn M (1990) *Qualitative Evaluation and Research Methods* (Second Edition). London: Sage Publications.

Pawson R (2006a) *Evidence-based policy: A realist perspective*. London: Sage Publications.

Pawson R (2006b) Digging for nuggets: how 'bad' research can yield 'good' evidence. *International Journal of Social Research Methodology* **9**: 127-142.

Pawson R, Boaz A, Grayson L, Long A, Barnes C (2003) *Types and Quality of Knowledge in Social Care*. London: Social Care Institute for Excellence.

Peers IS, Leadwith F, Johnston M (1993) *Community Youth Project on HIV/AIDS: Draft final report to health education authority*. Manchester: University of Manchester.

Peersman G (1996) *A Descriptive Mapping of Health Promotion Studies in Young People*. London: EPI-Centre, Social Science Research Unit.

Peersman G, Oakley A, Oliver S, Thomas J (1996) *Review of Effectiveness of Sexual Health Promotion Interventions for Young People*. London: EPI-Centre, Social Science Research Unit.

Peersman G, Oliver S (1997) *EPI-Centre Keywording Strategy: Data Collection for the BiblioMap Database*. London: EPI-Centre, Social Science Research Unit.

Peersman G, Oliver S, Oakley A (1997) *EPI-Centre Review Guidelines: Data Collection for the EPIC Database*. London: EPI-Centre.

Peersman G, Harden A, Oliver S (1998) *A Review of the Effectiveness of Workplace Health Promotion Interventions*. London: Health Education Authority.

Peersman G, Harden A, Oliver S (1999) *Effectiveness Reviews in Health Promotion*. London: EPPI-Centre, Social Science Research Unit.

Peersman G, Oliver S, Oakley A (2001) Systematic reviews of effectiveness. In: Oliver S, Peersman G (eds) *Using Research for Effective Health Promotion*. Buckingham: Open University.

Peterosino A, Turpin-Petrosino C, Buehler J (2003) Scared straight and other juvenile awareness programs for preventing juvenile delinquency: a systematic review of the randomized experimental evidence. *The ANNALS of the American Academy of Political and Social Science* **589**: 41-62.

Petticrew M, Roberts H (2006) *Systematic Reviews in the Social Sciences: A practical guide*. Oxford: Blackwell Publishing.

Pillemer DB (1984) Conceptual issues in research synthesis. *Journal of Special Education* **18**: 27-40.

Popay J (2005) Moving beyond floccinaucinihilipilification: enhancing the utility of systematic reviews. *Journal of Clinical Epidemiology* **58**:1079-80.

Popay J (2006) *Moving Beyond Effectiveness in Evidence Synthesis*. London: National Institute for Health and Clinical Excellence

Popay J, Rogers A, Williams G (1998) Rationale and standards for the systematic review of qualitative literature in health services research. *Qualitative Health Research* **8**:341-351.

Popay J, Arai L, Roberts H, Roen K (2003) *Preventing accidents in children – how can we improve our understanding of what really works?* London: Health Development Agency.

Popay J, Roberts H, Sowden A, Petticrew M, Aari L, Roen K, Rodgers M (2006) *Guidance on the Conduct of Narrative Synthesis in Systematic Reviews: A product*

from the ESRC Methods Programme. Lancaster: Institute for Health Research, Lancaster University.

Rees R, Harden A, Shepherd J, Brunton G, Oliver S, Oakley A (2001) *Young People and Physical Activity: A systematic review of barriers and facilitators*. London: EPPI-Centre, Social Science Research Unit, Institute of Education, University of London.

Rees R, Kavanagh J, Burchett H, Shepherd J, Brunton G, Harden A, Thomas J, Oliver S, Oakley A (2004) *HIV Health Promotion and Men who have Sex with Men (MSM): A systematic review of research relevant to the development and implementation of effective and appropriate interventions*. London: EPPI-Centre, Social Science Research Unit, Institute of Education, University of London.

Rees R, Kavanagh J, Harden A, Shepherd J, Brunton G, Oliver S, Oakley A (2006) Young people and physical activity: a systematic review matching their views to effective interventions. *Health Education Research* **21**: 806-825.

Reeves D (1999) *Deaf Gay Men, Safer Sex and HIV.* London: Gay Men Fighting AIDS.

Reichardt and Cook T (1979) Beyond qualitative versus quantitative methods. In: Cook T, Reichardt C (Eds.) *Qualitative and Quantitative Methods in Evaluation Research*. Beverly Hills, California: Sage Publications.

Reinharz S (1984) *On Becoming a Social Scientist*. Brunswick, NJ: Transaction Books.

Richie ND, Stenroos D, Getty A (1990) Using peer educators for a classroom-based AIDS program. *Journal of the American College of Health* **39**: 96-99.

Roberts C, Sibbald B (1998) Understanding controlled trials: randomising groups of patients. *British Medical Journal* **316**:1898-1900

Roberts I (2000) Randomized trials or the test of time? The story of human albumin administration. *Evaluation and Research in Education* **14**: 231-236.

Roberts K, Jones D, Fitzpatrick R, Dixon-Woods M, Abrams K (1999) Meta-analysis of qualitative and quantitative study evidence. *Proceedings of the 7th Cochrane Colloquium*. Universita S.Tommaso D'Aquino, Rome. Milan: Centro Cochrane Italiano.

Roberts SJ, McGuiness PJ, Bilton RF, Maxwell SM (1999) Dieting behavior among 11-15-year-old girls in Merseyside and the Northwest of England. *Journal of Adolescent Health* **25**: 62-67.

Rogers A, Adamson JE, McCarthy M (1997) Variations in health behaviours among inner city 12-year-olds from four ethnic groups. *Ethnicity and Health* **2**: 309-316.

Rorty R (1982) *Consequences of Pragmatism*. Minneapolis: University of Minnesota Press.

Rose H (1994) *Love, Power and Knowledge: towards a feminist transformation of the sciences*. London: Policy Press.

Rosenthal R (1991) *Meta-analytic Procedures for Social Research*. Thousand Oaks, California: Sage Publications.

Ross S (1995) 'Do I really have to eat that?': a qualitative study of schoolchildren's food choices and preferences. *Health Education Journal* **54**: 312-321.

Sackett D, Rosenberg W, Gray M, Haynes B, Richardson W (1996) Evidence-based medicine: what it is and what it isn't. *British Medical Journal* 312: 71-72.

Sandelowski M, Docherty S, Emden C (1997) Qualitative meta-synthesis: issues and techniques *Research in Nursing and Health* **20**: 365-371.

Sandelowski M, Barroso J (2002) Reading qualitative studies. *International Journal of Qualitative Methods* **1**: 1-47.

Sandelowski M, Barroso J (2003a) Creating metasummaries of qualitative findings. *Nursing Research* **52**: 226-233.

Sandelowski M, Barroso J (2003b) Classifying the findings in qualitative research. *Qualitative Health Research* **13**: 905-923.

Sandelowski M, Barroso J (2007) *Handbook for Synthesising Qualitative Research.* New York: Springer.

Schonbach K (1995) *Health Promotion and Peer Involvement for Youth.* Berlin: Themen & Konzepte.

Schwandt T (1998) The interpretative review of educational matters: is there any other kind? *Review of Educational Research* **68**:409-412.

Scott Porter Research and Marketing Ltd (2000) *Young people and Mental Well being*. Edinburgh: Health Education for Scotland.

Seale C (1999) Quality in qualitative research. *Qualitative Inquiry* **5**:465-478.

Seale C (2004) *Qualitative Research in Practice*. London: Sage.

Shadish W, Cook T, Campbell D (2002) *Experimental and Quasi-experimental Design.* Boston: Houghton-Mifflin.

Shepherd J, Peersman G, Weston R, Napuli I (2000) Cervical cancer and sexual lifestyle: a systematic review of health education interventions targeted at women. *Health Education Research* **15:** 681-694.

Shepherd J, Harden A, Rees R, Brunton G, Oliver S, Oakley A (2001) *Young People and Healthy Eating: A systematic review of barriers and facilitators.* London: EPPI-Centre, Social Science Research Unit, Institute of Education, University of London.

Shepherd J, Harden A, Rees R, Brunton G, Garcia J, Oliver S, Oakley A (2006) Young people and healthy eating: a systematic review of research on barriers and facilitators. *Health Education Research* **21**: 239-257.

Schulz K, Chalmers I, Hayes R, Altman D (1995) Empirical evidence of bias: dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *Journal of the American Medical Association* **273**:408-412.

Silverman D (2001) *Interpreting Qualitative Data* (2nd edition). London: Sage Publications.

Slavin R (1995) Best evidence synthesis: an intelligent alternative to meta-analysis. *Journal of Clinical Epidemiology* **48**: 9-18.

Smith JK (1984) The problem of criteria for judging interpretive inquiry. *Educational Evaluation and Policy Analysis* **6**:379-91.

Smith JK (1993) *After the Demise of Empiricism: The problem of judging social and educational inquiry*. Norwood, New Jersey: Ablex Publishing.

Social Exclusion Unit (2001) *Preventing Social Exclusion.* London: The Stationery Office.

Sparkes A (1998) Validity in qualitative inquiry and the problem of criteria: implications for sports psychology. *Sports Psychologist* **12**: 363-386.

Speller V, Learmonth A, Harrison D (1997) The search for evidence of effective health promotion. *British Medical Journal* **315**: 361 – 363.

Spencer L, Ritchie J, Lewis J, Dillon L (2003) *Quality in Qualitative Evaluation: A framework for assessing research evidence.* London: Cabinet Office.

Sports Council for Wales (2004) *A Matter of Fun and Games: Children's participation in Sport.* Cardiff: Sports Council for Wales.

Stephenson J, Strange V, Forrest S, Oakley A, Copas A, Allen E, Black S, Ali M, Monteiro H, Johnson A, The RIPPLE Study Team (2004) Pupil-led sex education in England (RIPPLE study): cluster-randomised intervention trial. *Lancet* **364**: 338-346.

Stock W (1994) Systematic coding for research synthesis. In: Cooper H, Hedges L (eds) *The Handbook of Research Synthesis.* New York: Russell Sage Foundation.

Strouse JS, Krajewski LA, Gilin SM (1990) Utilizing undergraduate students as peer discussion facilitators in human sexuality classes. *Journal of Sex Education and Therapy* **16**: 227-235.

Svenson GR (1998) *European Guidelines for Youth AIDS Peer Education*. Sweden: European Commission.

Sykes W (1990) Validity and reliability in qualitative market research: a review of the literature. *Journal of the Market Research Society* **32**: 289-328.

Tashakkori A, Teddlie C (1998) *Mixed Methodology: Combining qualitative and quantitative approaches*. Thousand Oaks, CA: sage Publications.

Thomas J (2002) *EPPI-Reviewer© 2.0 (Web edition)*. London: EPPI-Centre, Social Science Research Unit, Institute of Education, University of London.

Thomas J, Harden A (2003) Practical systems for systematic reviews of research to inform policy and practice in education. In: Anderson L, Bennett N (eds) *Evidence-Informed Policy and Practice in Educational Leadership and Management: Applications and controversies*. Paul Chapman Publishing.

Thomas J, Sutcliffe K, Harden A, Oakley A, Oliver S, Rees R, Brunton G, Kavanagh J (2003) *Children and Healthy Eating: A systematic review of barriers and facilitators*. London: EPPI-Centre, Social Science Research Unit, Institute of Education, University of London.

Thomas J, Harden A, Oakley A, Oliver S, Sutcliffe K, Rees R, Brunton G, Kavanagh J (2004) Integrating qualitative research with trials in systematic reviews: an example from public health. *British Medical Journal* **328**: 1010-12.

Thomas J, Sutcliffe K, Harden A, Oakley A, Oliver S, Rees R, Brunton G, Kavanagh J (2005) The barriers to, and the facilitators of, healthy eating among children: findings from a systematic review. In Cameron N, Norgan N, Ellison GTH (eds) *Childhood Obesity: Contemporary issues.* New York: CRC press

Thorne S, Paterson B, Acorn S, Canam C, Joachim G, Jillings C (2002) Chronic illness experience: insights from a metastudy. *Qualitative Health Research* **12**: 437-452.

Thorne S, Jensen L, Kearney M, Noblit G, Sandelowski M (2004) Qualitative metasynthesis: reflections on methodological orientation and ideological agenda. *Qualitative Health Research* **14**: 1342-1365.

Tilston C, Gregson K, Neale R (1991) Dietary awareness of primary school children. *British Food Journal* **93**: 25-29.

Tishelman C, Forss A, Sachs L, Lundgren E, Widmark C, Tornberg S (1999) Research on risk and risk in research: Theoretical and practical experiences from a multidisciplinary study on cervical cancer screening in urban sweden. *Qualitative Health Research* **9**: 45-60.

Tolley E, Girma M, Stanton-Wharmby A, Spate A, Milburn J (1998) *Young Opinions*. London: National Children's Bureau.

Tones K, Tilford S (1994) *Health Education, Effectiveness, Efficiency and Equity*. (Second edition). London: Chapman Hall.

Torrance H (2004) 'Quality in qualitative evaluation': a (very) critical response. *Building Research Capacity* **8**: 8-10.

Treloar C, Champness S, Simpson P, Higginbotham N (2000) Critical appraisal checklist for qualitative research studies. *Indian Journal of Pediatrics* **67**: 347-351.

Tones K, Tilford S (1994) *Health Education: Effectiveness, Efficiency and Equity*. London: Chapman & Hall.

Torgeson D, Roland M (1998) Understanding controlled trials: what is Zelen's design? *British Medical Journal* 316: 606.

Tuxworth B (1997) *The St Edmundsbury Fitness Survey 1993-1995*. Bury St Edmunds: St Edmundsbury Borough Council.

Vermeire E, Van Royen P, Griffiths F, Coenen S, Peremens L, Hendrickx K (2003) The critical appraisal of focus group research articles. *European Journal of General Practice* **8**: 104-108.

Vidich A, Lyman S (1994) Qualitative methods: their history in sociology and anthropology. In: Denzin N, Lincoln Y (Eds.) *Handbook of Qualitative Research.* London: Sage Publications.

Vulliamy G, Webb R (2001) The social construction of school exclusion rates: implications for evaluation methodology. *Educational Studies* **27**: 357-370.

Wallace A, Croucher K, Quilgars D, Baldwin S (2004) Meeting the challenge: developing systematic reviewing in social policy. *Policy and Politics* **32**: 455-470.

Wallace W (1971) *The Logic of Science in Sociology*. Chicago: Aldine-Atherton.

Warburton S (1998) Catch 'em young... Fit for Life Project... NT/3M National Nursing Awards. *Nursing Times* **94**: 46-47.

Ward M (2002) *Count Us In.* Edinburgh: Healthy Gay Scotland.

Watt RG, Sheiham A (1996) Dietary patterns and changes in inner city adolescents. *Journal of Human Nutrition and Dietetics* **9:** 451-461.

Watt RG, Sheiham A (1997) Towards an understanding of young people's conceptualisation of food and eating. *Health Education Journal* **56**: 340-349.

Weiss C (1979) The many meanings of research utilization. *Public Administration Review* **39**: 426-431.

Weston R (1998) *The Mythmakers in Health Promotion: Is the randomised controlled trial the gold standard?* Unpublished PhD Thesis, Southampton: University of Southampton.

Weston R, Harden A, Oakley A (1999) A systematic analysis of process evaluations [conference presentation]. *Proceedings of the 7th Cochrane Colloquium*. Universita S.Tommaso D'Aquino, Rome. Milan: Centro Cochrane Italiano.

Westwood M, Whiting P, Kleijnen J (2005) How does study quality affect the results of a diagnostic meta-analysis? *BMC Medical Research Methodology* **5**: 20

Wistow R, Schneider J (2003) Users' views on supported employment and social inclusion: a qualitative study of 30 people in work. *British Journal of Learning Disabilities* **31**:166-173.

White H (1994) Scientific communication and literature retrieval. In: Cooper H, Hedges L (eds) *The Handbook of Research Synthesis*. New York: Russell Sage Foundation.

Whittemore R, Chase S, Mandle C (2001) Validity in qualitative research. *Qualitative Health Research* **11**: 522-537.

Willig C (2001) *Introducing Qualitative Methods in Psychology: Adventures in theory and method*. Milton Keynes: Open University Press.

Wilton T, Keeble S, Doyal L, Walsh A, University of West England, South and West Regional Health Authority (1995) *The Effectiveness of Peer Education in Health Promotion: Theory and practice*. Bristol: University of the West of England.

Woolgar S (1991) *Knowledge and Reflexivity: New frontiers in the sociology of knowledge*. London: Sage Publications.

Wray K (2002) The epistemic significance of collaborative research. *Philosophy of Science* **69**: 150-168.

Wortman P (1994) Judging research quality. In: Cooper H, Hedges L (eds) *The Handbook of Research Synthesis*. New York: Russell Sage Foundation.

Yin R (2000) Rival explanations as an alternative to reforms as experiments. In Bickman L (ed.) *Validity and Social Experimentation: Donald Campbell's legacy*. Thousand Oaks, California: Sage Publications.

Yoneg O, Stwein L (1988) Reliability and validity: misnomers for qualitative research. *Canadian Journal of Nursing Research* **20**: 61-67.

EPI-CENTRE

Centre for Evaluation of Health Promotion and Social Interventions

REVIEW GUIDELINES

Data collection for the EPIC database

These guidelines have been developed by :

Greet Peersman, Sandy Oliver and Ann Oakley[1]

EPI-Centre, Social Science Research Unit (SSRU), Institute of Education, University of London, 18 Woburn Square, London WC1H 0NS, UK

Tel : +44 171 612 6393 Fax : +44 171 612 6400 E-mail : health@ioe.ac.uk

Note :

We have used the term **'review'** to indicate the standardised systematic data extraction from an evaluation study (published or unpublished).

The review results for each individual study are then combined using the **EPIC database**[2] to obtain an overview of the findings of the evaluation research in a particular health area, or by intervention type, intervention setting, intervention provider etc.

[1]building on the work of earlier grant holders (A. Oakley, D. Fullerton and J. Holland of the SSRU).

[2]the initial technical development of the EPIC database was carried out by D. Fullerton (SSRU) and P. Robertson (Institute of Education Computing Service) and is currently being continued by J. Thomas (EPI-Centre).

The **EPI-Centre** has built on previous work funded or commissioned by the Economic and Social Research Council, the Health Education Authority, the Medical Research Council, the NHS Centre for Reviews and Dissemination, the North East Thames Regional Health Authority, and the North Thames Regional Health Authority. The **EPI-Centre** is currently funded by the Department of Health for a specific programme of work on evidence-based health promotion.

EPI-CENTRE

Centre for Evaluation of Health Promotion and Social Interventions

REVIEW GUIDELINES

Data collection for the EPIC database

General comments

1. These review guidelines are for **intervention studies** only. In intervention studies the researcher attempts to change people's experience or situations by, for example, exposing people to an education programme, a skills training, or the use of a particular service, or by carrying out environmental modifications (eg improving housing conditions). Usually, but not always, the report will include an evaluation of the intervention, either a process evaluation or an outcome evaluation, ideally both :

- a **process evaluation** examines the acceptability and feasibility of an intervention, studies the ways in which the intervention is delivered, assesses the quality of the procedures performed by the programme staff etc. It is designed to describe what goes on rather than to establish whether or not the programme achieves its objectives, and may suggest ways in which the programme design and implementation could be improved.

- an **outcome evaluation** is designed to establish whether an intervention works or not, whether or not the intervention changes the outcomes (eg knowledge, attitudes, intentions, behaviour, service use) specified in the aims of the study.

2. The following pages help you record step-by-step the information required to assess the quality of a process and an outcome evaluation :

- **A.** How can the report be identified?
- **B.** Support for the study
- **C.** Type of study
- **D.** Description of the intervention
- **E.** Description of the study population
- **F.** Planning and process measures
- **G.** Quality of the outcome evaluation

This format allows the results of your review to be included in the EPIC database.

Instructions: **READ THE REPORT NOW**

Return to the report and **collect the required data** from it by **systematically completing the appropriate sections** that follow.

Choosing terms which best describe a research report and its findings : The aim of this work is to **apply terms systematically** to reports to enable the **inputting** of information and, subsequently, **to search and retrieve** the information **efficiently**. The choice of terms is necessarily restricted and occasionally the only term which usefully answers a question will be **"other", in which case the reviewer should add any specifications**. Reviewers should bear in mind that whenever they choose **"other",** searching and retrieving the detailed information that they include to explain **"other",** will be **severely limited**.

English spelling should be used throughout.

EPI-Centre, Social Science Research Unit, London University Institute of Education,
18 Woburn Square, London WC1H 0NS.

Reviewing process:

1 Name of reviewer	
2 Date of review	

A/How can the report be identified?

A1 Bibliography number *Fill in the number assigned to this paper for BiblioMap, the EPI-Centre bibliographic register. If this report is linked to (an)other report(s) of the same study, write in the bibliographic register number(s) of the linked report(s).*	..... linked to:..... .....
A2 Review number *Fill in the number assigned to this paper for EPIC, the EPI-Centre review database. If this report is linked to (an)other report(s) of the same study, write in the EPIC review number(s) of the linked report(s).*	..... linked to:..... .....
A3 How was this report found on this occasion? **Name of person who found this report:** ..... **If found by an electronic database search, indicate reference to search strategy used:**	1. Electronic database • a) AIDSLINE • b) Australian/British Education Index • c) CABhealth • d) CINAHL • e) Cochrane Library • f) EMBASE • g) ERIC • h) Health Planning • i) HealthPromis • j) MEDLINE • k) PsychINFO • l) PsycLIT • m) SIGLE • n) Social Science Citation Index • o) UnCover • p) other database *Specify*..... 2. Handsearch 3. Referenced in another report 4. Personal contact 5. Unknown 6. Other *Specify...:*.....
A4 Authors *Write in the name(s) of the author(s) (initials followed by surname for all)*	

A5a Title *Is this an article/ a chapter /a book/ a report? Circle which and write in the full title.*

A5b Language of report : English/other *Circle which and specify*.....
A5c Is it published/ in press/ unpublished? *Circle which*
A5d Date of report/ publication date *Write in*.....

A6 Journal/ book *Write in the full title of the journal or book and provide further details in A7a or A7b as appropriate*

A7a Details of journal	Volume number:..... Issue number *or N/A...*:..... Page numbers:.....
A7b Details of book	Place of publication:..... Name of publisher:..... Editor(s) of book:..... ..... Relevant page numbers:.....
A8 Abstract / Summary of key findings	1. Not included 2. Included
A9 Keywords	1. Not included 2. Included *List*:..... ....

A10 EPI-Centre Keywords	*List*:..... ....

B/Support for the study

B1 Name and address for correspondence *Write in details*	....
B2 Source of funding	1. Not stated 2. Stated *Write in details*..... ....

C/Type of study

- A **process evaluation** examines the acceptability and feasibility of an intervention, studies the ways in which the intervention is delivered, assesses the quality of the procedures performed by the programme staff etc. It is designed to describe what goes on rather than to establish whether or not the programme achieves its objectives, and may suggest ways in which the programme design and implementation could be improved.

- An **outcome evaluation** is designed to establish whether an intervention works or not, whether or not the intervention changes the outcomes (eg knowledge, attitudes, intentions, behaviour, service use) specified in the aims of the study.

- A **retrospective study** looks back in time.

- A **prospective study** looks forward in time.

C1 What type of study does this report describe?

Circle yes or no for each study type, and follow the instructions accordingly.

1.	Process evaluation	YES/ NO
2.	Outcome evaluation	YES/ NO
3.	Retrospective study	YES/ NO
4.	Prospective study	YES/ NO

Instructions:

If it is a **prospective process evaluation**, complete sections D,E,F

If it is a **prospective outcome evaluation**, complete sections D,E,G

If it is both a **prospective process and outcome evaluation,** complete sections D,E,F,G

If it is a **retrospective process/outcome evaluation**, complete sections D,E

D/Description of the intervention

D1 Country *Write in the country/ countries in which the intervention was carried out. NB This is not necessarily the same as the country of the research institution. If the study is conducted in more than one country, indicate them all;* ***for any part of Australia*** *- write Australia;* ***for any part of the USA*** *- write USA;* ***for any part of the UK*** *- write UK.*	....
D2 Topic area of the intervention *Circle* ***AS MANY AS APPROPRIATE****:* *To ensure optimal search and retrieval of information from the database, it is important to circle* ***ALL*** *those topic areas covered by the intervention in the* ***reviewer's judgement***	1. Accidents 2. Cancer - skin cancer 3. Cancer -other 4. Cardiovascular disease 5. Child neglect 6. Emotional abuse 7. Health inequalities 8. Mental health - eating disorder 9. Mental health - other 10. Nutrition - healthy eating 11. Obesity 12. Oral health 13. Physical abuse 14. Physical activity 15. Problem behaviour - delinquency 16. Problem behaviour - other 17. Sexual abuse 18. Sexual health - pregnancy prevention 19. Sexual health - STDs* including HIV/AIDS 20. Sexual health - other 21. Substance abuse - alcohol 22. Substance abuse - illicit drugs 23. Substance abuse - solvents 24. Substance abuse - tobacco 25. Other S*pecify*..... ** STDs: sexually transmitted diseases*
D3 Name of the programme *Write in the name of the programme if it is specified,* ***using the exact spelling as in the report*** *- this is the only time non-English spelling may be used.*	....

D4 Content of the intervention package *Describe the intervention* ***in detail****, whenever possible copying the authors' description from the report word for word (continue overleaf if more space is needed).* ***If specified in the report, also describe in detail what the control / comparison group(s) were exposed to****.*

D5 Aim(s) of the intervention

1. Not stated
2. Not explicitly stated *Write in, as worded by the reviewer*
3. Stated W*rite in, as stated by the authors. Continue overleaf if more space is needed*

D6 Year intervention started *Write in*	.....

D7 Theoretical model : *Some interventions are developed from theories/models of individual or group behaviour:*

Community orientated models:
Models which attempt to change attitudes or norms of a distinct group (eg prostitutes), by targeting a large proportion of the group. The intervention may involve self-efficacy or traditional education presentations, but also involves changing the context in which individuals operate by instilling new norms in all or most of the community members, relying on peer support and not on ***self****-efficacy.*

Health belief model:
This states that the likelihood of an individual adopting preventative behaviour(s) is dependent on four personal perceptions: their SUSCEPTIBILITY to the condition; the SERIOUSNESS of the condition; the BENEFITS and the efficacy of the preventative behaviour(s), and the extent of the BARRIERS to the behaviour(s).

Traditional education/Reasoned action models:
Models which assume that information presented to individuals will be absorbed directly, improving knowledge, or affecting their attitudes or behaviour. The objective is to alter knowledge only, although assumptions may be made about knowledge affecting behaviour. Models assume that individuals always act in a rational, logical way. Progressive media (e.g. video, theatre) may be used but information is still given in a didactic way.

Learning theory:
Two paradigms of learning are included under this heading: 1) respondent (or classical) conditioning, thought to account for the acquisition of a range of emotional and affective behaviour (such as phobias, anxiety, sexual dysfunction); key therapeutic interventions include graded in-vivo exposure and systemic desensitisation; and 2) operant (or instrumental) conditioning, which highlights the impact of environmental stimuli on behaviour; key therapeutic interventions include differential reinforcement, extinction, time-out, and punishment.

Social learning theory (socio-psychological/social cognitive/empowerment/self-esteem/self-efficacy etc.)*:*
This adds cognitive and observational learning to the respondent and operant paradigms (see above) and essentially says that human beings do not respond to stimuli, but interpret them. The key intervention derived from this theory is modelling (eg skills training).

Cognitive theories:
These emphasise the causal role of cognition in the development of behaviour, including problem behaviours. Interventions derived from these theories (Rational-Emotive Therapy; Cognitive Therapy; Stress Inoculation Therapy; Anger Control) focus therapeutic effort of effecting changes in the way people think (eg selective perception, misattribution, faulty information processing).

Psycho-dynamic theories:
These derive from the work of Freud, and stress the importance of early life experiences on the development of personality, particularly the psycho-social dramas and conflicts of key stages such as Oedipal phase.

Systems theory:
This emphasises the inter-connectedness of different parts of a whole, functioning entity such as the family, and conceptualises the problems experienced by individual family members as symptomatic of system 'malfunctioning'. Often problems are thought to arise because the family system has failed to re-establish an equilibrium following a system-disrupting crisis. Therapeutic strategies are aimed at assisting the family's return to a state of equilibrium and include : joining, reframing, and prescribing tasks.

Eco-behavioural/Ecological action models:
These focus attention on the influence of social factors, such as external stressors (eg poverty, serious life events), societal values, and developmental factors, and examine these within the framework of theories of learning.

D7a Theoretical model (as stated by the authors) *Indicate **ALL** the models which the **authors state** they have used in **the design of the intervention**.*	1. Not stated 2. Unclear 3. Community-orientated Model 4. Cognitive Theory 5. Eco-behavioural/Ecological Action Model 6. Health Belief Model 7. Learning Theory 8. Psycho-dynamic Theory 9. Social Learning Theory 10. Systems Theory 11. Traditional education/Reasoned Action Model 12. Other *Specify*.....
D7b Theoretical model (as judged by the reviewers) *Use the definitions provided to classify **the design of the intervention** within a theoretical model*	1. Unclear 2. Community-orientated Model 3. Cognitive Theory 4. Eco-behavioural/Ecological Action Model 5. Health Belief Model 6. Learning Theory 7. Psycho-dynamic Theory 8. Social Learning Theory 9. Systems Theory 10. Traditional education/Reasoned Action Model 11. Other *Specify*.....
D8 Intervention site *Circle **AS MANY AS APPROPRIATE***	1. Not stated. 2. Unclear 3. a) Community *Specify*..... b) Correctional institution *specify*..... 4. Day care centre 5. Educational institution *if not further specified* 5a) Educational institution - pre-school 5b) Educational institution - primary education 5c) Educational institution - secondary education 5d) Educational institution - tertiary education 6. Family centre 7. Health care unit *if not further specified* 7a) Health care unit - primary care 7b) Health care unit - hospital 7c) Health care unit - specialist clinic 8. Home 9. Hospice 10. Outreach 11. Residential care 12. Workplace *Specify*..... 13. Other S*pecify*.....
***12** : If workplace-based ONLY tick workplace ONLY and specify in as much detail as possible*	

D9 Length of the intervention *Choose the relevant category and* ***write in*** *the* ***exact*** *intervention length if specified in the report.* *When the* ***intervention is ongoing****,tick 14 and indicate the length of intervention as the length of the outcome assessment period*	1. Not stated 2. Not applicable 3. Unclear 4. One day or less *Specify*..... 5. 1day to 1 week *Specify*..... 6. 1 week to 1 month *Specify*..... 7. 1 to 3 months *Specify*..... 8. 3 to 6 months *Specify*..... 9. Up to 1 year *Specify*..... 10. Up to 2 years *Specify*..... 11. Up to 3 years *Specify*..... 12. 3 to 5 years *Specify*..... 13. more than five years *Specify*..... 14. Other S*pecify*.....
D10 Type of intervention *Circle* ***AS MANY AS APPROPRIATE*** **5.** ***Bio-feedback*** *i.e. feedback to an individual their* ***biological*** *measure(s) and/or behavioural/social risk status indicating potential health consequences e.g. the level of carbon monoxide in the blood related to smoking; cholesterol level related to cardiovascular disease; HIV-positive test related to AIDS; dietary fat intake;* **16.** ***Risk assesment*** *refers to the establishment of a risk profile (not solely relying on medical screening-see below-) for a particular adverse outcome, which is not fed back on an individual basis* **17.** ***Screening*** *refers to medical screening eg breast screening, ultrasound*	1. Not stated 2. Unclear 3. Advice/counselling 4. Anger management 5. Bio-feedback 6. Brief therapy 7. Casework 8. Environmental modification *Specify*..... 9. Family therapy 10. a) Increased access to resources *Specify*..... 10. b) Increased access to services *Specify*..... 11. Information/education 12. Legislation/regulation 13. a) Parent training b) Professional training 14. Physical activity 15. Practical skill development S*pecify*..... 16. Risk assessment 17. Screening 18. Social support 19. Other *Specify*.....
D11 Medium of intervention *Circle* ***AS MANY AS APPROPRIATE***	1. Not stated 2. Unclear 3. Curriculum materials 4. Discussion group session(s) 5. Incentives 6. Mass media *Specify*..... 7. One-to-one communication 8. Outreach 9. Practising practical skill 10. Presentation/ lecture 11. Printed materials/ posters 12. Role play 13. a)Theatre 13 b)Film/video/slides S*pecify*..... 14. Other *Specify*.....

D12 Person providing the intervention *Circle **AS MANY AS APPROPRIATE***	1. Not stated 2. Unclear 3. Not relevant (eg mass media) 4. a) Community 4. b) Community worker 5. Counsellor 6. Health professional S*pecify*..... 7. Health promotion/ education practitioner 8. Lay therapist 9. a) Parent 9. b) Peer S*pecify*..... 10. Psychologist 11. Researcher 12. Residential worker 13. Social worker 14. Teacher/ lecturer 15. Other *Specify*.....
D13 Number of people recruited <u>to provide</u> the intervention (and comparison condition) *(eg teachers or health professionals)*	1. Not stated *2.* Unclear *3.* Reported *Write in the numbers for the providers involved in the intervention and comparison groups as appropriate* ....

D14 How were the people <u>providing</u> the intervention recruited? *Write in. **Also give information on the providers involved in the comparison group(s) as appropriate**. Continue overleaf if more space is needed*

1. Not stated
2. Stated *Write in*

D15 Was special training given to people <u>providing</u> the intervention? *Provide as much detail as possible*	1. Not stated 2. Unclear 3. Yes *Specify*..... 4. No
D16 Did the authors indicate any costs related to the intervention? *Provide as much detail as possible*	1. Yes *Specify*..... 2. No

E/Description of the study population

E1 Characteristics of the study population at point of entry in the study *Circle **AS MANY AS APPROPRIATE***	
***A. Age group**: Record age range and numbers/proportion of the population in each age group if specified. We define children (0-10 years); young people (11-21 years); adults (22-54 years); and older people (55+ years).*	**A. Age group** 1. Not stated 2. Children *age & nos./%*..... 3. Young people *age & nos./%*..... 4. Adults *age & nos./%*..... 5. Older people *age & nos./%*..... 6. General population/ Mixed (no information)
***B. Definition of class: Working class** describes people employed in **manual work**. **Middle class** describes employed in **non-manual work**. Alternatively, information about **tenure** (renting or owner occupier) or **age of leaving full time education** or **eligibility for financial benefits** may be useful indicators of social class. Record numbers/proportion of population in each class if specified*	**B. Class** 1. Not stated 2. Working class *nos/%*..... 3. Middle class *nos/ %*..... 4. Authors' description *Write in*..... ..... **C. Ethnicity** *Write in authors' quantitative and qualitative description* ....
***D. Region:** Record numbers/proportion of population in each type of region if specified*	**D. Region** 1. Not stated 2. Rural *nos/%*..... 3. Urban *nos/%*.....
***E. Sex:** Record numbers/proportion of population of each sex if specified*	**E. Sex** 1. Not stated 2. Female 3. Male 4. Mixed sex *nos/%*.....
***F.Sexual orientation:** Record numbers/proportion of population with each orientation if specified.*	**F. Sexual orientation** 1. Not stated 2. Heterosexual *nos/%*..... 3. Homosexual *nos/%*..... 4. Bisexual *nos/%*.....
G.Family:**Answer **AS MANY AS APPROPRIATE	**G. Information about the family** 1. Not stated 2. Family size *Specify*..... 3. Family structure *Specify*..... 4. Housing conditions *Specify*.....

H. Other information *Continue overleaf if more space is needed*

.....

.....

E2 Sampling and Recruitment procedures *Write in. Continue overleaf if more space is needed*

E3 Were any incentives provided to recruit people into the study?	1. Not stated 2. Yes *Specify* .....
E4 Were participants asked for their consent before entering the study? *This refers to the **eligible sample***	1. No/Not stated 2. Unclear 3. Not relevant (eg mass media) 4. Requested from participants *State which participants* .... 5. Requested from others *State which others*..... ....
E5 What proportion of the sample consented to the study?	1. Not stated 2. Unclear 3. Not relevant (eg mass media) 4. Stated *Write in*..... .....
E6 Were views on research priorities sought? *If the authors report the views of the population in the study (ie study population), tick (3). If they report the views of a similar population, although not directly involved in this study (ie target population) tick (3). If they report the views of others, such as those providing the intervention tick (4) and specify which others.*	1. Unclear 2. No/Not stated 3. Yes, from study/target population 4. Yes, from others S*pecify*..... ....
E7 Were views on the intervention sought? *If the authors report the views of the population in the study (ie study population), tick (3). If they report the views of a similar population, although not directly involved in this study (ie target population) tick (3). If they report the views of others, such as those providing the intervention tick (4) and specify which others.*	1. Unclear 2. No/Not stated 3. Yes, from study/target population 4. Yes, from others *Specify*..... ....

E8 Were views on the evaluation sought? *If the authors report the views of the population in the study (ie study population), tick (3). If they report the views of a similar population, although not directly involved in this study (ie target population) tick (3). If they report the views of others, such as those providing the intervention tick (4) and specify which others.*	1. Unclear 2. No/Not stated 3. Yes, from study/target population 4. Yes, from others S*pecify*..... ....

F/Planning and process measures

This section is organised in four sub-sections :

1. Development of the Intervention
2. Development of the Process/Outcome Evaluation
3. The Process Evaluation
4. Dissemination and Implementation

Complete ALL sections except for **Section 3** which **only** needs to be completed in case the report describes a **process evaluation** (cfr. answer to C1.1 is Yes)

1. Development of the Intervention

F1 Was the intervention based on a needs assessment?

4 : comparative need *is derived from examining for example the services provided in one area to one population and using this as the basis to determine the sort of services needed in another area with a similar population*

5 : expressed need *refers to what one can infer about the need of a community by observing their use of services*

6 : felt need *is what people say they want or what they think are the problems that need addressing*

7 : normative need *refers to what expert opinion defines as need*

Specify further where possible

1. Not stated
2. Yes, no further information provided/information unclear
3. Yes, reference to source of further information given *Write in*

.....

4. Yes, based on 'comparative need'
5. Yes, based on 'expressed need'
6. Yes, based on 'felt need'
7. Yes, based on 'normative need'
8. Yes, other *Specify*

.....

9. No, but there was another rationale for delivering this intervention/undertaking this study *Specify*

.....

.....

.....

F2 Who identified the aim(s) of the intervention?

Specify further where possible

1. Not stated
2. Unclear
3. Evaluator
4. Funder
5. Health promotion practitioner
6. Intervention provider
7. (A sample of the) study population *Specify*

.....

8. (A sample of the) target population *Specify*

.....

9. Other *Specify*

.....

F3 Who was involved in the development of the intervention? *Specify further where possible*	1. Not stated 2. Unclear 3. Evaluator 4. Funder 5. Health promotion practitioner 6. Intervention provider 7. (A sample of the) study population *Specify* ..... 8. (A sample of the) target population *Specify* ..... 9. Other *Specify* .....
F4 Was the intervention piloted? *A pilot study involves preliminary use of some or all of the elements of the intervention in order to refine the intervention or its delivery. This does not include similar interventions tested by* ***others.*** *Specify further where possible*	1. Not stated 2. Unclear 3. The authors consider this study to be a pilot 4. Yes, previously piloted with the study population 5. Yes, previously piloted with a sample of the target population *Specify* ..... 6. Yes, previously piloted with others *Specify* ..... 7. No
F5 Do the authors indicate any barriers to developing/delivering the intervention?	*1.* Yes *Write in* .... 2. No
F6 Do the authors indicate any factors favourable to developing/delivering the intervention?	*1.* Yes *Write in* .... 2. No

2. Development of the Process/Outcome Evaluation	
F7 Were views on the evaluation design sought? *Specify further where possible*	1. Not stated 2. Unclear 3. Yes, from the funder 4. Yes, from a health promotion practitioner 5. Yes, from the intervention provider 6. Yes, from the study population 7. Yes, from a sample of the target population *Specify* ..... 8. Yes, from others *Specify* ..... 9. No
F8 Who identified the range of processes/outcomes to be addressed? *Specify further where possible*	1. Not stated 2. Unclear 3. Evaluator 4. Funder 5. Health promotion practitioner 6. Intervention provider 7. (A sample of the) study population *Specify* ..... 8. (A sample of the) target population *Specify* ..... 9. Other *Specify* .....
F9 Who carried out the evaluation? *Specify further where possible*	1. Not stated 2. Unclear 3. Health promotion practitioner 4. Researcher *Specify* ..... 5. (Individuals from the) target population *Specify* ..... 6. Other *Specify* .....
F10 Does the report describe how the evaluators were selected? *Specify further where possible*	1. No 2. Unclear 3. Yes *Specify* .....

F11 Was special training provided for the evaluators? *Specify further where possible*	1. Not stated 2. Unclear 3. Yes *Specify* ..... 4. No

3. The Process Evaluation

Instruction : Ignore this Section if the report does **not** describe a **process evaluation**

F12 Which processes were evaluated? *Circle **AS MANY AS APPROPRIATE*** *Specify further where possible*	1. Acceptability of the intervention 2. Accessibility of the intervention/programme reach 3. Consultation/collaboration/partnerships *Specify* ..... 4. Content of the intervention 5. a) Implementation of the intervention b) Costs associated with the intervention 6. Management and responsibility 7. Quality of the programme materials 8. Skills and training of the intervention providers 9. Other *Specify* .....
F13 What methods were used to collect data on the processes involved? *Circle **AS MANY AS APPROPRIATE*** *Specify further where possible*	1. Not stated 2. Unclear 3. Documentation 4. Focus group 5. Interview 6. Observation 7. Self-completion report or diary/questionnaire 8. Other *Specify* .....
F14 Who was the data collected from? *Specify further where possible*	1. Not stated 2. Unclear 3. Intervention provider W*rite in nrs* ..... 4. (A sample of the) study population W*rite in nrs* ..... 5. Other *Specify and write in nrs* .....

F15 When did the evaluation take place in relation to the intervention? *Circle **AS MANY AS APPROPRIATE*** *Specify further where possible*	1. Not stated 2. Unclear 3. Afterwards *Specify* ..... 4. Concurrently 5. For a limited period during the intervention 6. Other *Specify* .....
F16 About which processes do the authors offer conclusions? *Circle **AS MANY AS APPROPRIATE*** *Write in **ALL** Conclusions*	1. None 2. Unclear 3. Acceptability of the intervention 4. Accessibility of the intervention/programme reach 5. Consultation/collaboration/partnerships 6. Content of the intervention 7. a) Implementation of the intervention b) Costs associated with the intervention 8. Management and responsibility 9. Quality of the programme materials 10. Skills and training of the intervention providers 11. Other *Specify* .....
F17 Are there any obvious inconsistencies in the reporting of the evaluation data?	*1.* Yes *Write in* ..... 2. No
F18 Do the data presented substantiate the authors' findings? *Specify further where possible*	1. Unclear 2. Yes 3. No *Specify* .....
4. Dissemination and Implementation	
F19 Who were the findings reported back to? *Specify further where possible*	1. Not stated 2. Unclear 3. Yes, to all in the study population *4.* Yes, to some in the study population *Specify* ..... 5. Yes, to all intervention providers *6.* Yes, to some intervention providers *Specify* ..... 7. Yes, to the target population 8. Other *Specify* .....

G/Quality of the outcome evaluation

G1 What were the aims of the evaluation? *Circle **ONE ONLY***	1. 2. 3. To evaluate a single intervention 4. To compare different intensities/levels of an intervention 5. To evaluate the generalisability of an intervention 6. To compare different interventions 7. Other *Specify*.....

G2 STUDY DESIGNS

Post test only: *a group receives an intervention, and outcomes are measured after the intervention only*

Pre- and post-test: *a group receives an intervention and outcomes are measured both before and after the intervention*

A trial: *compares groups receiving different interventions or different intensities/levels of an intervention with each other; and/or with a group which does not receive any intervention at all.*

G2 What was the design of the evaluation? *In the **reviewer's judgement,** using above definitions*

1. Post-test only
2. Pre- and post-test
3. Trial
4. Other *(specify)*.....

Reviewers may find it helpful to draw a flow diagram overleaf to depict the evaluation design and include all relevant numbers in the different groups before attempting to answer the following questions.

G3 What proportion of the eligible population were recruited?	1. 100% 2. 80% > 100% 3. 60% > 80% 4. 40% > 60% 5. 20% > 40% 6. 10% > 20% 7. 5% > 10% 8. 2% > 5% 9. 0 > 2% 10. Not stated 11. Not relevant 12. Participation not voluntary
G4 Number of participants in each intervention and control/comparison group (on the basis of those from whom baseline data were collected) or **for the study population as a whole if only one group**	1. Not stated 2. Unclear 3. Reported *Write in number for each group* ....

G5 How were participants allocated to intervention and control/comparison groups?	1. Not relevant (study not a trial) 2. Not stated 3. Unclear 4. Non random *Write in*..... 5. Random, no information given 6. Random, information given *Write in*..... .....
G6 What was the unit of allocation into each intervention and control/ comparison group?	1. Not relevant (study not a trial) 2. Not stated 3. Unclear 4. Community 5. Family 6. Group/Class 7. Individuals 8. Institution 9. Region 10. Other *Specify*.....
G7 Was the allocation to intervention and control/comparison groups done blind?	1. Not relevant (study not a trial) 2. Not stated 3. Unclear 4. Yes 5. No
G8 Were participants aware which group (intervention/control/comparison) **they were in for the evaluation?**	1. Not relevant (study not a trial) 2. Not stated 3. Unclear 4. Yes 5. No
G9 Was outcome measurement done blind? *(ie were those assessing the outcomes aware whether the participant had been in a control/comparison or intervention group?)*	1. Not relevant (study not a trial) 2. Not stated 3. Unclear 4. Yes 5. No
) How did the different group(s) in the trial compare to one another ***(in the reviewer's judgement)*** *ups are likely to be* ***equivalent*** *if they are drawn from similar populations and have* ***similar demographic variables and pre-test outcome measures***	1. Not relevant (study not a trial) 2. Unclear 3. Equivalent 4. Non-equivalent 5. Other S*pecify*..... .....
G11 Was information on socio-demographic variables reported before the intervention began?	1. No 2. Unclear 3. Information for some individuals/groups only) Information only for those individuals remaining in the study 4. Information for all individuals/groups (or for study population as a whole if only one group) 5. Information for the study population in general 6. Other *Specify* .....

G12 What outcomes did the authors say they were <u>intending</u> to measure (ie as described in the aims of the evaluation)? *Circle **AS MANY AS POSSIBLE** and **Specify** where possible* ***8** : as determined by a clinical test eg blood pressure, cholesterol level* ***9** : including anxiety, depression, other mental health state; other examples : pregnancy, coronary heart disease*	1. Not stated 2. Unclear 3. Access to/availability of resources *Specify*..... 4. Attitudes *Specify*..... 5. Awareness/Beliefs *Specify*..... 6. Behaviour (observed) *Specify*..... 7. Behaviour (reported) *Specify*..... 8. Clinical risk factor(s) *Specify*..... 9. Health problem or state (prevalence and/or incidence) *Specify*..... 10. Intentions *Specify*..... 11. Knowledge *Specify*..... 12. Legislation/regulation *Specify*..... 13. Practical skill *Specify*..... 14. Self-efficacy/self-esteem/self-confidence *Specify* ..... 15. Service use *Specify*..... 16. Other *Specify*..... .....
G13 Were data on outcome variables reported before the intervention began?	1. No 2. Unclear 3. a) Information for some individuals/groups only *Specify*..... 3. b) Information only for those individuals remaining in the study 4. Information for all individuals/groups (or for study population as a whole if only one group) 5. Information for the study population in general 6. Information for some outcomes only *Specify* ..... 7. No baseline data reported, only change reported 8. Other *Specify*..... .....
G14a What was the attrition or participation rate? (on the basis of those from whom baseline data were collected) *Make clear whether it is attrition or participation that is reported*	1. Not stated 2. Unclear 3. Not relevant (eg mass media) 4. Reported for the study population as a whole *Write in* ..... 5. Reported for one/some group(s) *Write in* ..... 6. Reported for all groups (or for study population as a whole if only one group) *Write in* ....

G14b Was any information provided on those who dropped out of the study?	1. Unclear 2. Not relevant (eg mass media) 3. Yes, reported (write in) 4. No
G15 What sort of measurement tool(s) is/are used to collect outcome data?	1. Not stated 2. Unclear 3. Clinical test 4. Interview 5. Observation 6. Practical test 7. Psychological test 8. Self-completion report or diary/questionnaire 9. Other S*pecify*..... ....
G16 Name(s) of measurement tool(s)	1. Not stated 2. Stated *Write in*..... ....
G17 Has the measurement tool been used in a previous published study?	1. Unclear 2. No 3. Yes *Specify source*..... .....
G18 Were data on outcome variables reported after the intervention? *Compare the outcomes reported with your answers in G12* ***4*** *: ie all those remaining in the study*	1. No 2. Unclear 3. Information for some individuals/groups only *Specify* ..... 4. Information for all individuals/groups (or population as a whole if one group) 5. Information for the study population in general 6. Information for some outcomes only *Specify* ..... 7. No final data reported, only change reported 8. Other *Specify*.....
G19 Number of outcome assessment periods *ie how many times were data on outcome variables collected after the intervention?*	1. Not stated 2. Unclear 3. One 4. Two 5. Three 6. Four or more
G20 Timing(s) of pre-intervention measurement(s)	1. Not stated 2. Unclear 3. Stated *Write in*..... 4. Not relevant (ie no pre-intervention measurement)

G21 Timing(s) of post-intervention measurement(s) *Choose one of the categories and indicate the* ***exact timings*** *if specified in the report* *! **Beware** : the option "**immediately after intervention**" is at the bottom of the list!*	1. Not stated 2. Unclear 3. Up to 1 month 4. Up to 3 months 5. 3 to 6 months 6. 6 to 12 months 7. 1 to 2 years 8. 2 to 3 years 9. 3 to 5 years 10. More than 5 years 11. None 12. Immediately after intervention
G22 Data analysis method: ***"Intention to treat"*** *means that data were analysed on the basis of the original number of participants recruited into the different groups.* ***"Treatment received"*** *means data were analysed on the basis of the number of participants remaining in the groups at the time of measurement.*	1. Not relevant (study not a trial) 2. Not stated 3. Unclear 4. 'Intention to Treat' 5. 'Treatment Received'
23 Unit of data analysis *Were the results reported according to the unit of allocation? For example, if individual people were allocated to different groups, results from individuals should be analysed and reported; whereas if schools were allocated to different groups, results from each school should be analysed and reported.*	1. Not relevant (study not a trial) 2. Not stated 3. Unclear 4. Same as unit of allocation 5. Different from unit of allocation *Specify* ....
G24 Are there any obvious errors in the numerical reporting? W*rite in*	1. No 2. Yes (write in)
G25 Impact of the intervention *Compare outcomes reported with your answers in G12*	1. Unclear 2. Not stated 3. Reported for some outcomes only *Specify* .... 4. Reported for all outcomes

G26 Is the study replicable from this report?

A study is replicable if

1. *there is a clear description of the design of the evaluation (G1-23 give clear answers)*
2. *there is a clear description of the intervention content (D4,D5,D10)*
3. *there is clear description of how the intervention was delivered (D11,12,14)*

Give answers to each of the sections 1,2 and 3:

1a The evaluation design is replicable
1b The evaluation design is not replicable
1c Reference to source of further information given (design)

2a The intervention content is replicable
2b The intervention content is not replicable
2c Reference to source of further information given (content)

3a The intervention delivery is replicable
3b The intervention delivery is not replicable
3c Reference to source of further information given (delivery)

G27 Is the outcome evaluation sound?

The outcome evaluation is sound if:

- *it has an equivalent control or comparison group, not necessarily randomised (G10.3)*
- *it reports pre-intervention data for all individuals/groups (G13.4) An exception is made for studies using the Solomon Four Group design in which intervention and control/ comparison groups are further randomised to receive pre-intervention surveys or not (ie pre-intervention data are not available for half the subjects in the intervention and control/ comparison groups).*
- *it reports post-intervention data for all individuals/groups (G18.4)*
- *it reports on all outcome measures as described in the aims of the study (G25.4)*

If this report is linked to other reports (A1 and A2), a decision on soundness of the evaluation should be made on the basis of all reports. Use section G32 to clarify your decision.

1. Sound
2. Not sound
3. Reviewer judges study sound despite discrepancy with 4 quality criteria

Clarify.....

.....

.....

.....

.....

.....

G28 Is the outcome evaluation "gold standard"?

An outcome evaluation is "gold standard" if ***in addition to being sound*** *it:*

- *has clearly defined aims (D5.3))*
- *describes the intervention and the evaluation design well enough for both to be replicated (G26.1a or 1c* ***and*** *2a or 2c)*
- *uses random allocation to different groups, including to the control/comparison group (G5.5 or 6)*
- *reports numbers of people assigned to each intervention and control/comparison group (G4.3)*
- *reports the attrition rates for each intervention and control/comparison group (G14a.6)*

If this report is linked to other reports (A1 and A2), a decision on gold standard status of the evaluation should be made on the basis of all reports. Use section G32 to clarify your decision.

1. Gold standard
2. Not gold standard
3. Reviewer judges gold standard despite discrepancy with 8 quality criteria

Clarify.....

.....

.....

.....

.....

.....

G30 Note: *Conclusions of **sound studies using control group(s)** (ie no intervention) should be discussed separately from the conclusions of **sound studies using comparison group(s)** (ie different intervention or different intensity/level(s) of an intervention). When reporting conclusions on the effect of an intervention, that has been evaluated **without the use of a control group,** it is **not** possible to provide **absolute measures of effects**, but only **relative measures**. In the case of **sound** studies including **control group(s)**, the conclusion may be that the intervention is effective, ineffective etc.; in the case of **sound** studies including **comparison group(s)** only, the conclusion may be that the intervention is more/less effective, etc.*

G31 Was there agreement between the authors and the reviewer?	1. Yes 2. No

G32 If the reviewer <u>disagrees</u> with the authors about the effect of the intervention, give the reasons here.

Even if the authors and reviewers **agree** about effect there may be **other issues** for example about **generalisability** or **sample sizes** which are worth commenting on here. *Continue overleaf if more space is needed*

G33 Please record any studies listed in the bibliography which may be relevant (methodology studies, outcome evaluations or process evaluations). *Continue overleaf if more space is needed*

G34 Did this report describe a randomised controlled trial? *The study is a randomised controlled trial if you ticked G2.3 **and** G5.5 or G5.6*	1. Yes 2. No

APPENDIX B: Strategy used in study two to collect data from tools to assess the quality of 'qualitative' research

Section A: Identifying details and reason for inclusion in review

A.1 Name of reviewer *Person who has completed this data extraction*	A.1.1 Details
A.2 Date of review	A.2.1 Details
A.3 Author(s) of tool *Indicate all author names*	A.3.1 Details
A.4 How many people authored the tool	A.4.1 One A.4.2 More than one
A.5 Country(ies) in which author(s) is/are based	A.5.1 UK A.5.2 Other European country *(please specify)* A.5.3 USA A.5.4 Other country *please specify*
A.6 Discipline or professional background of author(s) *It may help to indicate the universtiy and department in which the author(s) is(are) based.*	A.6.1 Education A.6.2 Health sciences - nursing A.6.3 Health sciences - other *please specify* A.6.4 Psychology A.6.5 Sociology A.6.6 Other (please specify)
A.7 Multidisciplinary team	A.7.1 Yes A.7.2 No
A.8 Year in which tool was published/reported	A.8.1 Before 1980 A.8.2 1980 A.8.3 1981 A.8.4 1982 A.8.5 1983 A.8.6 1984 A.8.7 1985 A.8.8 1986 A.8.9 1987

	A.8.10 1988 A.8.11 1989 A.8.12 1990 A.8.13 1991 A.8.14 1992 A.8.15 1993 A.8.16 1994 A.8.17 1995 A.8.18 1996 A.8.19 1997 A.8.20 1998 A.8.21 1999 A.8.22 2000 A.8.23 2001 A.8.24 2002 A.8.25 2003
A.9 Format in which tool is published/reported *i.e. as a jounral article, within a book*	A.9.1 Journal article A.9.2 Book A.9.3 'Stand alone' report A.9.4 Other (please specify)
A.10 Does/do the author(s) explicitly state that they have produced a tool for assessing the quality of a 'qualitative' study from its written report?	A.10.1 Yes, developed for use in systematic reviews A.10.2 Yes, developed for use by journal editors or peer referees A.10.3 Yes, developed for others (please specify) A.10.4 No
A.11 If the author(s) does not/do not state that they have produced a tool for readers to use to assess the quality of a 'qualitative' study from its written report, please state why the reviewer has included this report in this review of tools.	A.11.1 Details A.11.2 Not applicable
A.12 Was there any funding for tool developement?	A.12.1 Yes A.12.2 No/Not stated

Section B: Conceptual underpinnings and development of tool

B.1 What definition of 'qualitative' research is given/used by the author(s)?	B.1.1 Explicitly stated B.1.2 Implicitly stated B.1.3 Not stated/unclear
B.2 Where does/do the author(s) locate themselves in terms of the debates about assessing the quality of 'qualitative' research?	B.2.1 Details
B.3 What reasons does/do the author(s) give for why they developed the tool?	B.3.1 Explicitly stated B.3.2 Implicitly stated B.3.3 Not stated/unclear
B.4 How was the tool developed?	B.4.1 Explicitly stated B.4.2 Implicitly stated B.4.3 Not stated/unclear
B.5 Is there any other information given by the author(s) concerning the conceptual underpinnings or development of the tool?	B.5.1 Yes *please specify* B.5.2 No
B.6 Manifesto	B.6.1 Yes B.6.2 No

Section C: Description of tool

C.1 Is the tool intended for use with a particular type of 'qualitative' research?	C.1.1 Yes *please specify the type of 'qualitative' research the tool is intended for.*

C.2 Is the tool intended for use within a particular discipline/applied field of study?	C.2.1 Yes, Education C.2.2 Yes, Health care C.2.3 Yes, Social care C.2.4 Yes, Psychology C.2.5 Yes, Sociology C.2.6 Yes, Other (please specify) C.2.7 No *please specify*
C.3 What aspects of quality does (do) the author(s) say the tool assesses? *Try to describe the standards that the authors say the tool is trying to assess e.g. validity, credibility.*	C.3.1 Details
C.4 Describe the tool including its format and layout *List the items in the tool and any associated guidance*	C.4.1 Details
C.5 How many items (e.g. questions, statement of a particular standard) are listed in the tool for reviewers to answer/judge the study against?	C.5.1 Details
C.6 Does/do the author(s) provide guidance for how reviewers should make a judgement on the items in the tool?	C.6.1 Yes C.6.2 No C.6.3 Yes, for all items in the tool C.6.4 Yes, for some items in the tool only
C.7 Does/do the author(s) provide a structured answer format for reviewers to record their judgement on the items in the tool? *e.g. yes/no categories*	C.7.1 Yes C.7.2 No C.7.3 Yes, for all items in the tool C.7.4 Yes, for some items in the tool
C.8 Does/do the author(s) offer guidance for how a reviewer should use the tool to make an overall judgement on the quality of a particular study? *e.g. calculate an overall score*	C.8.1 Yes *please specify* C.8.2 No *please specify*
C.9 How many items within the tool direct the reviewer to assess	C.9.1 Theoretical or empirical

aspects of the study in the following areas? *Fill in as many as apply (NB: try not to place any one item in more than one category)* *Please give an indication of what the items in each area are trying to assess such as: presence or absence of a particular procedure (e.g. was X carried out/ obtained?); quality of reporting (e.g. if the reviewer is asked to assess whether some aspect of the study is clearly stated or adequately described etc); or quality of implementation (e.g. if the reviewer is asked to assess whether some aspect of the study (other than reporting) is appropriate, adequate, sufficient).*	framework of study C.9.2 Aims, research questions and phenomenon under study C.9.3 Design C.9.4 Setting of study C.9.5 Sampling and sample *Please indicate which items are to do with sampling and which are to do with the actual sample obtained* C.9.6 Sample C.9.7 Data collection C.9.8 Data analysis C.9.9 Findings *i.e the products of data analysis (presentation and substance)* C.9.10 Relevance *e.g. generalisability, implications of findings* C.9.11 Ethical issues C.9.12 Other aspect of study (please specify)
C.10 Does the tool refer to any of the following techniques advocated for increasing rigour in 'qualitative' research? *Please indicate whether the tool contains specific items about these or whether they are just referred to in the guidance.*	C.10.1 Analytic induction C.10.2 Audit trail C.10.3 Coding by more than one researcher C.10.4 Constant comparison C.10.5 Deviant/negative case analysis C.10.6 Grounded theory C.10.7 Low-inference descriptors C.10.8 Prolonged engagement C.10.9 Purposive/theoretical sampling C.10.10 Reflection C.10.11 Respondent validation C.10.12 Standardised methods for collection and/or transcribing data C.10.13 Thick description

	C.10.14 Triangulation C.10.15 Training, qualifications, or characteristics of researcher C.10.16 Other (please specify)

Section D: Evaluation of tool

D.1 Has the tool been tested and if so, what were the results?	D.1.1 Yes D.1.2 No
D.2 Who tested the tool?	D.2.1 The authors of the tool D.2.2 Others (please specify) D.2.3 Not applicable (i.e. tool has not been tested)
D.3 Please comment on the utility of this tool from a systematic review perspective	D.3.1 Details